2017

Using high throughput data to understand microbes and their interaction with the environment

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

AND

COLLEGE OF ENGINEERING

Dissertation

USING HIGH THROUGHPUT DATA TO UNDERSTAND MICROBES AND THEIR INTERACTION WITH THE ENVIRONMENT

by

AMRITA KAR

B.Tech., West Bengal University of Technology, 2010
M.S., Boston University, 2013

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2017

Approved by

First Reader _____

Daniel Segre, Ph.D.
Professor of Biology

Second Reader _____

Kirill Korolev, Ph.D.
Associate Professor of Physics

Third Reader _____

James Galagan, Ph.D.
Associate Professor of Biomedical Engineering and Microbiology

DEDICATION

To Anita, Suddhabrata and Saurav

For inspiring and guiding me through the journey of life

USING HIGH THROUGHPUT DATA TO UNDERSTAND MICROBES AND THEIR INTERACTION WITH THE ENVIRONMENT

AMRITA KAR

Boston University Graduate School of Arts and Sciences

and

College of Engineering, 2017

Major Professor: Daniel Segre`, Professor of Biology

ABSTRACT

The human microbiome ecosystem plays numerous, yet poorly understood beneficial roles in human health. It can shape the immune response and provide essential vitamins and enzymes to the host. The different environments present in the human host are a major determinant of community composition. Conversely, the presence of certain bacteria in specific parts of the human body is sometimes associated with an increased chance of pathologies. Advances in DNA sequencing have increased our understanding of the relationship of microbes with the environment. However, sequencing data alone is unlikely to provide such understanding without the help of appropriate computational models and analyses.

For the first part of this thesis, I applied to the infant gut microbiome an approach previously used to understand the order of colonization of microbial biofilms. Available metagenomic sequencing data from infant fecal samples collected for 2.5 years was queried to test whether or not the gut colonization process is a multi-step process, in which the organisms that are prevalent at a given time are closely related, in their

metabolic capabilities, to the organisms present at the previous time step. I further used network expansion algorithms previously developed for the study of large-scale biogeochemical evolution, to explore the dynamics and diet-dependency of the gut microbiome. These analyses suggest that metabolic relatedness among organisms is an important factor in the colonization process.

The second part of my thesis explores the role of *H. pylori* in gastric cancer. I analyzed public microarray data for gastric AGS cancer cell lines infected with different strains of *H. pylori* differing in pathogenicity. Relative to uninfected AGS cell lines, low-pathogenic *H. pylori* strain displayed no major metabolic dysregulation, consistent with the fact that *H. pylori* does not cause inflammation/gastric cancer in a majority of the human population. However, gastric AGS cell lines infected with highly pathogenic strains showed more significant differences, including the upregulation of purine metabolism, possibly consistent with an inflammatory response.

The results in this dissertation thus offer insights into how the interplay between metabolic activity of human-associated microbes and their surrounding environment plays an important role in the colonization process as well as in pathogenesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACOA	Acetyl coenzyme A
ADP	Adenosine di phosphate
AMP	Adenosine mono phosphate
ATP	Adenosine tri phosphate
BLAST	Basic Local Alignment Search Tool
COA	Coenzyme A
COG	Clusters of Orthologous Groups
DNA	Deoxyribonucleic acid
EC	Enzyme Commission
FAD	Flavin adenine dinucleotide
GEO	Gene Expression Omnibus
GSEA	Gene Set Enrichment Analysis
HMO	Human Milk Oligosaccharides
KEGG	Kyoto Encyclopedia of Genes and Genomes
KS	Kolmogorov-Smirnov
MAC	Microbiota Accessible Carbohydrates
mRNA	Messenger ribonucleic acid
NADH	Nicotinamide adenine dinucleotide
NADP	Nicotinamide adenine dinucleotide phosphate
NADPH	Nicotinamide adenine dinucleotide phosphate
NCBI	National Center for Biotechnology Information

OTU	Operational Taxonomic Unit
PCA	Principal Component Analysis
Pfam	Protein Family
QIIME	Quantitative Insights into Microbial Ecology
SRA	Sequence Read Archive
TCA	Tricarboxylic acid
USDA	United States Department of Agriculture

CHAPTER ONE

Introduction

1.1 The Human Microbiome History and Predictive Modeling

Several individual microbes, as well as the whole ecosystem known as the human microbiome, play a fundamental role in different aspects of human health. They influence the rise of major diseases, like obesity, metabolic conditions, inflammatory bowel diseases, diabetes and even cancer. When they do not switch to a "disease state", however, human-associated microbial communities are not only harmless, but also symbiotically contributing important functions to the human host. For example, they are responsible for producing essential vitamins like B12 and are likely helpful in keeping pathogens away [1] [2]. The power of microbiome-based diagnostics and therapy has already inspired several researchers and commercial ventures, especially based on the established notion that microbiome transplants can cure certain disease conditions [3]. As a result, microbiome research has grown leaps and bounds within the recent years. Interestingly, most biomedical research in this area is based on empirical attempts and experimental methods. In addition to detecting microbial compositions before and after treatment many researchers have focused their attention on trying to understand the basic nature of microbial community interactions and spatio-temporal organization, as well as the interactions of the microbes with the host. In parallel to large experimental data collection and statistical analyses, these types of questions are also amenable to computer simulations and mechanistic models. These types of analyses can provide better understanding of the function and dynamics of the microbial community

and could also spearhead personalized microbiome-based interventions which can either be used as probiotics or biomarkers of diseases [4].

Computational models and analyses constitute an essential component of microbiome research, and can range from sequencing and analysis pipelines to simulations and machine learning methods. Broadly speaking, computational approaches, in addition to helping process and organize raw sequencing data, aim at characterizing or predicting different states of the microbiome and understanding their health-related consequences. The most active research area in computational biology of microbiomes is probably the development of sequence-based methods to optimally process amplicon (typically 16S rRNA) sequence and whole metagenomic data, and to use these data for functional and pathway mapping [5]. Such functional analyses of microbiome data typically involve mapping the sequences onto databeses containing detailed annotations on specific processes. Databases commonly used include KEGG, Metacyc, Pfam, Uniprot of COG database [6] [7] [8] [9]. Not less important are tools for data visualization, such as Cytoscape and VisANT, which can be used to analyze individual pathways, or more complex networks [10] [11]. Additional recent approaches to microbiome research involve Machine Learning algorithms, such as Supporting Vector Machines (SVMs) and neural networks. Training SVMs with algorithms can categorize the microbiome and help predict dysbiotic (imbalance of organisms in the gut) or non-dysbiotic stages [12]. In neural networks approaches, the microbial community structure and species abundances can be represented in matrix format, and used to predict the community behavior [13]. In other types of computational methods that can be thought of as

"reverse ecology" approaches network analyses have been used to identify sets of compounds that would be necessary for the growth of certain organisms [14]. Finally, as explored in this dissertation, it is possible to use "network expansion" algorithms to try and infer, from an initial set of metabolites (known as seed set, and potentially mimicking the diet) what other metabolites/reactions would likely appear in a given community [15].

1.2 The Human Microbiome Colonization

Understanding the dynamics and interactions of different microbial species is one of the most important challenges of human microbiome research, as this would pave the way to rational design of therapies. A specific aspect of the dynamics and interactions of human-associated microbial communities I will focus on is process of colonization. The order of colonization can be viewed as a temporally ordered sequence of taxa whose progression can be affected by several factors like the type of birth, diet, environment, age, stress, antibiotics and inflammation. The colonization of a new niche in the human host is an extremely complex process. In some cases, it is possible to identify in the colonization process (e.g. in the gut) two distinct steps. In the first step, 'early colonizers' first bind to the mucus membrane of the host, and prime the environment for 'late colonizers' [16]. The colonization process can depend strongly on the early colonizers, and give rise to a diverse set of possible healthy and stable communities. If the colonization process is disrupted at the initial phases, it can lead to serious repercussions for the set of organisms that would be dependent on the first set of microbes. For example, these secondary microbes could be commensal and beneficial organisms, whose absence could

cause the invasion of pathogens or disease [17]. The colonization process is also known to impact significantly the health state of the host's immune system [18]. In building computer models for the order of colonization of a microbial community, it is important to incorporate functional (metabolic) information about microbes as well as known details about interactions with the environment. Some of this complexity is captured efficiently by genome scale flux balance models of microbial metabolism. These approaches, initially developed to understand the physiology of individual organism, are now routinely applied to study complex communities[19]. Ultimately, the hope is that such mechanistic understanding of human-associated microbial communities will help us identify probiotics that can stir the microbiome towards desired states [20] and explore new avenues for the generation of new natural products, such as antimicrobials [21].

In addition to studying computationally the general properties of the colonization process of a microbiome, my dissertation work focuses on the possible health effects of the interactions between a specific bacterium (*H. pylori*) and the human host, especially in connection to its possible implication in gastric cancer.

1.3 *H. pylori* and its interaction with the environment

Helicobacter pylori is a unique organism that can thrive in the acidic human stomach lining. To survive in such a harsh environment, the bacterium converts urease in the cytoplasm into carbon dioxide and ammonia. This lowers the pH in the stomach and allows the bacteria to thrive. It survives best within pH of 4.0 – 8.0 [22]. *H.*

pylori is known to play an important part in gastric tumorogenisis by causing inflammation. This is believed to be mediated by the urease enzyme. The production of ammonia not only damages the stomach lining but causes dysregulations in several metabolic and signaling pathways (*Helicobacter pylori* infection, oncogenic pathways and epigenetic mechanisms in gastric carcinogenesis – Ding et al Future Oncol 2011). This microbe can be present in most people and cause no harm. It is becoming increasingly clear that only specific strains of this bacteria are associated with higher virulence to gastric cancer [23].

1.4 Dissertation Aims

My dissertation work has been centered around the following questions:

1. Is the colonization process of a human-associated microbiome a multi-step process dominated by metabolic relatedness of organisms?

2. Can one use diet information, and a network expansion algorithm REF to infer properties of a community and the dependence of its organisms on the specific molecular component of the diet?

3. Is mRNA dysregulation involved in *H. pylori* pathogenicity? In particular, do we observe dysregulation of pathways that can mediate interactions with the microbe? How do different strains exacerbate to different extents gastric cancer?

CHAPTER TWO
Exploring the role of inter-species metabolic distance and diet on the order of colonization in human gut microbiome

2.1 Background

The genetic information present in the human body can be viewed as the set of the genomes of several different microbial species in addition to the human genome itself. The large number of prokaryotic, viral, and fungal genomes present in the human host, known collectively as the human microbiome, constitutes an integral of the human host and contributes significantly to its health, through functions that are still poorly understood[24]. One of the many ways in which the human microbiome affects us is the complex set of interactions between microbes and the immune system [25]. Furthermore, the microbiome is known to provide enzymatic capabilities responsible for the biosynthesis of specific molecules that are necessary for our health, but that ore otherwise non-producible by our body. These metabolic functions include the production of vitamins (B12) as well as important enzymes that can degrade complex carbohydrates, such as CAZymes (carbohydrate active enzymes) [26] [27]. In addition to interacting with the human host, microbes are thought to interact heavily with each other, forming ecological networks that may significantly affect the microbiome composition and dynamics, as well its influence on our body. One of the ways in which the human environment and interspecies interactions can dramatically affect the whole microbiome dynamics is through the specific order of colonization, and associated structure of multi-species communities.

The colonization process of a microbial community, and its temporal and spatial aspects, have been studied extensively [28] [29]. In environmental research, microbial successions are known to consistently involve specific taxa with cyclical recurrence that may reflect seasons, or with specific orders of appearance after disturbances [30]. Microbial successions are also very important in the human microbiome. The idea of using sequencing to map the order of colonization of the infant gut had been used already in 2002 [31]. For the study of more directly accessible microbial colonies, such as in the case of the human oral microbiome, detailed scenarios of the order of colonization had been mapped through direct observations of the biofilm, or in vitro experiments of pairwise interactions, summarized in the literature as global putative networks [32] [16].

While in its simplest form, the order of colonization of a community can be viewed as a temporally ordered sequence of taxa, the seeding and progression of a microbial ecosystem can depend on multiple factors, and the existence of a clearly defined and deterministic succession is still debatable. One hypothesis about microbiome colonization in the human gut posits that it constitutes a complex process comprising of two major steps. In the first step, 'early colonizers' first bind to the host, and prime the mucus lining, preparing the ground for 'late colonizers' [16]. This process can create a diverse community, as suggested by the 'saturable niche hypothesis' [33]. According to this theory if a particular species has saturated or occupied a particular niche of the human intestine lining, it will prevent other strains of the same species to occupy the same area. It will however help related species to grow around it, reducing competition and promoting cooperation. This is because if there are too many similar species they would

be competing for the same food source. Conversely if they were functionally dependent they would facilitate each other. For example, *Bacteroides thetaiotaomicron* utilizes glycans from the host and the host diet to grow. Knocking out genes or transcription factors that target the glycan utilization pathway thus affects the glycan catabolism and colonization of other microbes [34]. This is because these specific glycans that are secreted into the extracellular space by this bacterium can be a food or energy source for many bacteria and thus promotes diversity and stability in a community [35].

This initial organization of the microbiome plays an important part in defense against pathogens and also prevention of colonic tumorogenesis [36]. Extensive research has been done to show how early colonizers interact with each other and with the late colonizers. For example, on the tooth surface the early colonizer *Streptococcus* has special enzymes that bind to the salivary receptors on the tooth surface and prime the environment for later colonizers like *Actinomyces* and *Fusobacterium*. This creates a dynamic community in the oral cavity and is important in avoiding periodontal diseases [37]. The process of early colonizers priming the niche for late colonizers is also seen in other microbial niches in the human host.

In humans, microbiome colonization can be affected by several non-pathologic factors like the type of birth – vaginal or caesarean, feeding patterns – breast fed or formula fed, diet at adult age and also the environment [38] [39]. However, disease and/or antibiotics can cause significant imbalance in the taxonomic abundance and functions of human-associated microbes, disrupting the ecosystem structure, and affecting the succession upon colonization or re-colonization. This disruption in microbiome structure

can in itself have further consequences, i.e. cause severe diseases due to abnormal immune response, lack of vitamins and disrupt metabolic pathways, e.g. those involving carbohydrate degradation [2 26 40]. Therefore, it is important to try and better understand the colonization process, its predictability, and its dependence on mechanisms whose manipulation may help control the microbiome for therapeutic purposes.

The microbial colonization process is often perceived as being strongly dependent on ecological interactions and host diet, such that organisms close to each other in the succession tend to be dependent on each other, or to have related metabolic functions [41]. In this chapter I will use computational analyses to ask whether and to what extent the colonization process in the human gut is recapitulated and explainable in terms of the metabolic enzymes and functions associated with the organisms that form the succession. In other words, will a random colonization be more robust than the actual colonization structure. A similar idea had been previously explored in the oral microbiome [42] taking advantage of the known structure of the dental plaque biofilm. In the oral microbiome case, it was found that the layers of microbial species could be translated into layers of enzymatic functions, which could be viewed as building gradually upon each other in a way that was consistent with the known order of colonization. For the human gut microbiome, the problem is somehow different, as the actual network of bacteria and their interactions is not known, and the information available is typically just the composition of the microbiome at different time points. Analyzing this information in search for signals of a metabolic basis for the order of colonization poses certain challenges, which will be described later in the chapter.

The dataset that I will be using is a time series 16S rRNA data that divides the samples into different diets [43] (See Section 2.2.1 for addtional details). To explore the order of colonization I will mainly be focusing on comparing the microbial composition between different time points within a given diet regime ("intra-diet"). This is because the hypothesis of the order of colonization algorithm we will apply is that organisms build on each other's capabilities, based on an otherwise steady environment. Thus, for each diet, we will explore a different colonization process.

The effect of diet on the microbiome, and on the colonization process has been studied abundantly. For example, prior work showed that shifting mice from a low fat, plant polysaccharide rich diet to a high fat, high sugar diet caused a shift in the microbial diversity [44]. Mice fed on complex microbiota-accessible carbohydrates (MACs) found in dietary fiber have a higher microbial diversity than those fed on simple carbohydrates. This diversity even decreases across generations and is irreversible unless missing genera over the generations are manually re-introduced [45].

Given the dependence of the microbiome on diet, it has been hypothesized that knowledge of the microbial community composition can provide direct information about diet. Eating a plant based diet causes increase in organisms like *Roseburia*, *Eubacterium rectale*, and *Ruminococcus bromii* that are robust in complex polysaccharide digestion whereas an animal based diet causes an increase in bile tolerant organisms like *Alistipes*, *Bilophila*, and *Bacteroides* that metabolize proteins [46].

Diet can affect host energetics thereby affecting various aspects of health and increasing the risk of obesity [47] [48]. Another crucial aspect of host-microbiome interaction is the

effect that diet can have on different microbial taxa, which in turn can modulate the immune system. For example, human breast milk contains many oligosaccharides (HMOs). Its microbiome is dominated by *Bifidobacterium infantis.* HMOs are one of the most abundant compounds found in breast milk and they are a major food source for *B. infantis*. This bacterium, in turn, modulates in an important way the baby's immune response [49].

In order to understand the shift in microbial community as a function of diet, several studies have performed community analyses based on sequencing data. However, very few attempts have been made towards understanding this connection from the perspective of microbial ecology and network-based approaches [50]. The working hypothesis of my project is that a network based approach can be valuable to relate diet and microbiome. Thus, in this chapter I will be discussing a mechanistic approach to look at dynamically changing communities in the infant gut as a function of diet. My analysis will be based on molecular data about the diet, network algorithms using KEGG pathway information, and 16S rRNA sequencing data. The diet information is represented in the form of an array of molecules or constituents that make up the food, with corresponding estimated abundance. This was carefully done utilizing the metadata collected for the subject and manually curated with the help of a Bioinformatics graduate students, Lili Ge and is discussed in details below.

2.2 Methods

2.2.1 Data

The data I used for the project was obtained from previously published work. I will summarize here briefly some key information about the dataset. Infant gut 16S rRNA sequencing data that was collected from fecal samples of a healthy infant from the time of birth to an age of 2.5 years during diaper changes [43]. As described in the original paper, around 60 fecal samples were collected including one from the mother at the time of birth. The average read length in the data per sample is about 250. Following sample collection, the microbial cells were lysed and purified the V2 region of the 16S rRNA genes was subjected to amplification using forward and reverse primers which were represented as the following 5′-*GCCTTGCCAGCCCGCTCAGTC***AGAGTTTGATCCTGGCTCAG**-3 and 5′-*GCCTCCCTCGCGCCATCAGNNNNNNNNNNNNCA*-**TGCTGCCTCCCGTAGGAGT**-3′ respectively. The italicized tag is 454 Life Sciences' primer and the bold tag is the bacterial primer. Following these steps, and appropriate barcoding, the samples underwent a 454 Pyrosequencing using a Roche 454 FLX pyrosequencer. The 16S sequencing reads were then submitted to the SRA (Sequence Read Archive) database from where I downloaded the fastq files for analysis [51]. There was also additional information collected in terms of the different diets and antibiotics administered to the infant. Based on this the time series data was divided into 4 steps. Table 1 shows the break up for the diets.

Step	Sample/Days	Number of Time-points	Diet
1	Meconium,4,5,6,10,14,16,19,23,27,31,33,48,55,57,63,64,70,77,84	20	Breast Milk
2	118,128,133,134,139,141,146,161	8	Breast Milk + Rice Cereal
3	172,173,195,202,206,240,244,252,265,273,280,294,297	13	Breast Milk + Rice Cereal + Formula
4	454,468,469,539,568,623,745,831,835,838	10	Cow's Milk + Adult Diet

Table 1: Description of the time series data and the diets administered

2.2.2 Pre-processing

I analyzed the 16S data using the QIIME (Quantitative Insights Into Microbial Ecology; qiime.org; version 1.9.0) software to gather information regarding what genera were present in the time course samples [52]. The QIIME pipeline was constructed in the following steps. First a mapping file was created that mapped the samples to their barcode (a DNA sequence assigned to the samples that allows them to be multiplexed and read), linkerprimer sequence (a sequence that helps to amplify the 16s rRNA in the sample) and also contains metadata like infant age and diets. Next, the BLAST program was used to analyze the results and align it to a reference database [53]. Following prior work with 16S rRNA, I used the Greengenes database to attribute samples to organisms. Using this database I could match the data samplings to 97% of the genome and find representative OTUs (operational taxonomic unit) [54]. Moreover, to avoid including spurious matches, I used a fairly stringent cutoff for the e-value in QIIME. In particular,

this was set to a high significance level (less than or equal to 0.001). Following these steps, taxonomy was assigned to the samples using a cut-off of 90% similarity to Greengenes using BLAST [53]. Phylogenetic tress were then created in QIIME using their default FastTree algorithm [55]. This led to the generation of a final OTU (operational taxonomic unit) table that contains the representative organisms is a particular time-point and their abundances.

2.2.3 Constructing a network of interconnected and random organisms for the infant gut

This OTU table with assigned taxonomy was then used to calculate if metabolism plays a key role in determining the order of colonization among the diets. To do so, I compared organisms appearing in an ordered progression to a random sampling of the data. The algorithm is best described through the following toy model example. Let there be 3 lists of organisms for 3 time points 1, 2 and 3 respectively. Assume that at each time point three different taxa are present, e.g. {A, B, C}, {A, D, E} and {E, F, G} respectively. For the order-preserving path I tried many strategies.

Strategy 1

I randomized the organism order appearing within a day and then found pairwise distances of all pairs. This meant my list of organisms for time point 1, 2 and 3 are now {B, A, C}, {D, E, A} and {G, F, E}. I then calculated the pairwise distance for {B-A, A-C, C-D, D-E, E-A, A-G, G-F, F-E} which represented my ordered path.

Strategy 2

I randomly chose a genus from one of the time points without allowing repeats and construct a network. This means that my new network is choosing one from each time point {A, D, F} and the pairwise distance of the path would be {A-D, D-F}.

Strategy 3: I randomly chose a genus from one of the time points allowing overlapping organisms in the time point if it happened to show up. The path created in the method could be {A, A, E} and the distances would be {A-A, A-E}.

Strategy 4

I also removed the low occurring genera and followed strategy 2 and 3 i.e. if genus A was less abundant it was removed before sampling.

Strategy 5

The last strategy that we chose was to a modification of Strategy 3 in which we look between diets. This is because sample collection had higher frequency between diets compared to across all time-points. Also, the organisms between diets might be not related closely with respect to their biochemical function and enzyme content.

For the random path let there be 3 lists for time point 1, 2 and 3 respectively given as {A, B, C}, {A, D, E} and {E, F, G}. Once again I experimented with various strategies.

Strategy 1

I shuffled all the days and organisms found within the infant. So, the list would now look like time 2 - {D, A, F}, time 3 – {A, G, E} and time 1 – {E, B, C}. The path would then be {D-A, A-F, F-A, A-G, G-E, E-E, E-B, B-C}.

Strategy 2

I shuffled only all the organisms without changing the order of days within it. So, the list would now look like time 1 - {D, A, E}, time 2 – {A, G, E} and time 3 – {F, B, C}. The path would then be {D-A, A-E, E-A, A-G, G-E, E-F, F-B, B -C}.

Strategy 3

Next I only randomized the days. So now the path would be time point 2 – {A, D, E}, time point 1 – {A, B, C} and time point 3 – {E, F, G}. The distance would be calculated between {A-D, D-E, E-A, A-B, B-C, C-E, E-F, F-G}.

Strategy 4

(A) For this the days and organisms were shuffled like Strategy 1 but I only chose 1 organism from every time point without repeat. Using the example from Strategy 1 the path would now be {D, G, B} and the distances would be {D-G, G-B}.

(B) I also tried this same strategy removing the less abundant organisms. So, if A is less abundant it was removed from the analysis before using Strategy 4A.

Strategy 5

The last strategy was the same as Strategy 4A except we only looked at organisms in the same diet. Repeats were allowed. The reason is the same as highlighted in Strategy 5 for ordered path selection.

Finally, I also wanted to compare the random path in a diet to a completely random set of organisms. For this I queried KEGG for a list of all genera and created an arbitrary list of those to compare them to the ordered and random diet [6].

2.2.4 Curating genus level enzyme information and Calculating enzyme distances for consecutive genera in a network

Each organism is associated with certain biochemical functions, which are encoded by enzymes. The KEGG database contains information for these enzymes in the form of Enzyme Commission (EC) number that are ascribed to an organism's genome [6]. Querying KEGG rest style API I can create a binary matrix by estimating the presence or absence of a particular enzyme in a species. This was done on July 22, 2016. I queried around 4300 organisms out of which 3815 belonged to bacteria, with a total of 5518 enzymes. Since I wanted to perform a genus-level analysis, but KEGG returned matching to species (with several species belonging to a given genus), I created a mapping file that mapped each species to its genus. This gave me 961 bacterial genera.

To create a binary genus to enzyme matrix the following strategy was adopted. If the enzyme was present in 50 percent of the species in the genera it is denoted as 1 (present) else, it's 0 (absent). Given this binary profile for each genus we can now calculate the jaccard's distance between the pairs of genera.

For example, consider two genera A and B with binary EC vectors V^A and V^B. If enzyme x is present in genus A , $V^A = 1$ and if it's absent, $V^A = 0$. Using this, Jaccard's Distance is calculated as follows:

$$\mathrm{J(A,B)} = 1 - \frac{\mathrm{V^A} \cap \mathrm{V^B}}{\mathrm{V^A} \cup \mathrm{V^B}}$$

It is the difference of the ratio of enzymes that are present in both A and B versus the total enzymes that are present in A and B. If J(A,B) = 1 then the genera are highly diverse and when J(A,B) = 0 that means that genus A and B are metabolically alike.

The jaccard's distance metric was used between pairs of genera in 3 pathways – the order preserving, the random path in a diet and the random path in KEGG to calculate the metabolic proximity of organisms. This process was repeated 1000 times for each of the 3 networks to increase the sampling space and calculate the efficiency in the order of colonization using enzyme data.

2.2.5 Curation of Diet for Network Expansion

The next part of the chapter deals with using network based approaches to understand the association of diet with the microbiome. For this information about the infant diet was duly noted in the Koenig et al paper. They classified the time series data into 4 steps based on the diet. Using this information, I and another graduate student Lili Ge were able to look at the metabolites that constituted the diet. Literature reviews were done for a detailed curation of the diet [56 57 58 59 60 61]. This was then analyzed with the help of the USDA website Release 28, version September 2015 [62]. Since the food intake changes with age, we even went a bit further and calculated the consumption of that nutrient per day based on the infant's intake of a food source during a certain age. The ages chosen and their diets are highlighted in Table 2. Also, highlighted in Supplementary Table 1, 2, 3 and 4 are the complete metabolite set of diets.

Diet	Age
Mother's Milk	2 months
Mother's Milk + Rice Cereal	4 months
Mother's Milk + Rice Cereal + Formula + Peas	6 months
Cow's Milk + Adult Diet (Fruits – Apples, Banana, Proteins – Boiled Egg, Chicken Nuggets, Fish Sticks, Lentils, Diary – Yogurt, Cheddar Cheese, Butter, Starch – Rice, Cereal, Pasta, Bread, Potatoes)	1.5 years

Table 2: Representative ages on the infant with the diet that was curated

There were several challenges that we had to address while curating this dataset. One of them was dealing with complex macromolecules. Some examples of these were casein found in dairy products and lutein a carotenoid found in green leafy vegetables and starch [63] [64] [65]. The challenge with them was that giant molecules could not take part in simple reactions even if they were important. To overcome this these complex macro-molecules were broken down into simpler amino acids using literature references that are utilized by the body [66] [67] [68]. One such example of this is casein, which is mentioned in Supplementary Table 5. Finally, not all the compounds in the diet was used for network expansion. This was because when most of the compounds were present in the diets the difference between the diets became minimal. As a result, several approaches were used.

Approach 1

This involved only keeping the abundant seed compounds. Different thresholds for the compounds were tried which were – 5%, 10%, 25%, 30%, 40%, 50%, 60%, 75% and 100% (all the compounds). However, since we were only taking the top compounds the

difference could not be noted as the top compounds did not change sufficiently over diets. Also, if I used all the seed set metabolites the network expansion of all the diets converged to the same unless the goal was looking at pre-adult diet vs adult diet.

Approach 2

The strategy that I used thus involved looking at log abundance. This is because the abundance of the compounds was measured over number of orders of magnitude. I used many cut-offs to see the network size and then decided to do a case study of all the metabolites (as used in Supplementary Table 1-4), median abundance of -5 and a strict cut off of 0. This metabolite set is highlighted Supplementary Table 6 and 7 and Supplementary Figure 1. Table 3 also shows the number of metabolites included for the seed set threshold.

Cut-off	Diet 1	Diet 2	Diet 3	Diet 4
All	59	65	74	129
-5	54	60	71	124
0	11	12	12	26

Table 3: Number of metabolites corresponding to the log abundance thresholds

2.2.6 Network Expansion and calculating enzyme profile for a diet

Network expansion is an iterative algorithm that can predict the state final state of system given a set of initial inputs (generally 2) known as the "seed set". This algorithm has been extensively used in in many evolutionary studies and also analyzing the input to

output relationship of a metabolic system over time [69] [70]. If we look at the diet aspect it does supply important molecules and affects the metabolic system and also if we look at the microbiome it does function as a list of organisms with enzymes. As a result, this approach can be used as a novel method to find the effect of diet on the microbiome and the 16S data can be used as a tool to verify this network approach. The seed set in this case would involve two things, firstly a set of all the reactions available from KEGG and secondly the list of metabolites curated from the diet above.

Reactions from KEGG were filtered based on several criteria for the input reaction list. Firstly, the reactions from KEGG involve several co-factors that might be important for the sustenance of a reaction. The assumption during this process was that the co-factors were presumed to be present and as a result their dependence on a reaction was eliminated from the enzyme stoichiometric matrix. These co-factors were ATP, ADP, AMP, NAD, NADH, NADP, NADPH, FAD, FADH, COA and ACOA. Secondly, a list of spontaneous reactions was added to the KEGG reactions that occur without any enzyme. Thirdly, unbalanced chemical reactions were removed from the list as they don't adhere to the law of conservation of mass. Apart from that I also obtained a list of compounds that correspond to reactions from KEGG. Thus, if we summarize the inputs from KEGG and the diet as the following matrix:

$$K = \{\text{compounds}_{=1\ldots5944}, r_{j=1\ldots6880}\}$$

$$\text{where } r_j \text{ is the set of reactions}$$

$$D_{i=1\ldots4} = \text{compounds present in each diet i}$$

In the final step, network expansion acts as an operator between the information from KEGG and the curated diet. As mentioned above, when I put in the initial seed set I get a subsequent list of compounds and reactions that can be generated from the iterative step and so on and so forth. Finally, after the network has reached saturation I stop the algorithm and it gives me a final list of compounds and reactions. The reactions were then mapped to their consequent enzymes and finally a binary matrix or binary profile α for each diet was created as:

$$\text{Network Expansion} \rightarrow \text{NE}$$

$$\text{NE}(D_i, K) \rightarrow \boldsymbol{\alpha_i} = [\mathbf{0}, \mathbf{1}, \ldots, \mathbf{0}, \mathbf{1}]$$

$$\text{where } \alpha_{i,j} = 1 \text{ if reaction } r_j \text{ is present in in diet i for } \text{NE}(D_i, K)$$

2.2.7 Calculating the Euclidean distance between diet profiles and the 16S data

The first step in this method was that I calculated the probability of a genus present in the infant microbiome data to have a particular enzyme. The genera that are found in the data was isolated and then multiplied with the abundance of those organisms that was calculated by QIIME and their enzyme content from KEGG. This gave me the enzyme abundance matrix.

The second step involved calculating the frequency of species that were found in a particular genus. The enzyme abundance matrix in the first step was then divided by the number of species found in a genus in the second step to get the enzyme probability of

the genera found in the data. Let us call this final matrix δ_g where g represents the genus.

2.2.8 Finding the top 20 organisms in the infant microbiome that can differentiate between diets 1, 2, 3 and 4

The network expansion profile for each diet is α_i and the enzyme profile each genus in the data is δ_g. The Euclidean distance between them is denoted as:

$$E(\alpha_i, \delta_g) = d_{i,j}$$

where d is the distance of genus g with the NE in diet i

The distances for each diet are then ranked with the lowest distance values. This would mean that those genera have the closest resemblance to the network expansion have smallest distance. The thing to note is that all the genera are present in the infant data but they might be ranked differently in the diets due to their enzymatic similarity with the expanded network. I chose to look at the top 20 organisms that resemble an expanded network expansion profile.

2.3 Results

2.3.1 Validating Results with Source Data

Since QIIME was also used in the source paper for the data with some minute changes in the method of assigning taxonomy to the infant specimens I wanted to see if our method could recapitulate similar results [43]. For this, I compared the phyla

abundances that were obtained in the in the published paper alongside abundances that we calculated with QIIME as seen in Table 4.

From Figure 1 we can see that there are 2 types of colonizers as mentioned before in the chapter. The 'early' colonizers are colored in purple. They are the Firmicutes. They are then followed by the 'late' colonizers *Bacteroidetes* in blue. The *Bacteroidetes* flourishing in the later half makes sense as they are mainly dependent on carbohydrates in the host diet [71]. Further validation can be seen in Table 4. Thus, my analysis is in agreement with the published analysis results. Apart from that, looking at Figure 1, the blast analysis integrated with QIIME yields more phyla compared to the analysis performed. This was because the source data used cd-hit which uses a short word filtering [72]. For a sample to be similar to a reference genome with a particular threshold they have to have a certain number of base pairs overlapping. This way cd-hit can skip many pairwise alignments which a typical blast search would not.

A

Organisms (Koenig)	**Diet 1**	**Diet 2**	**Diet 3**	**Diet 4**
Actinobacteria	1.5 %	2 %	-	-
Bacteroidetes	0.5 %	-	50 %	50 %
Proteobacteria	10 %	20 %	19 %	8.6 %
Firmicutes	88 %	78 %	31 %	41.4 %

B

Organisms (QIIME)	Diet 1	Diet 2	Diet 3	Diet 4
Actinobacteria	0.4 %	2 %	0.5 %	0.2 %
Bacteroidetes	0.1 %	0.1 %	48 %	60 %
Proteobacteria	15 %	15.9 %	14 %	11 %
Firmicutes	80 %	82 %	35 %	28.7 %
Others	-	-	2.5 %	0.1 %

Table 4: (A) Percent of organisms from Koenig paper, (B) Percent of organisms from the QIIME pipeline

2.3.2 Order of Colonization is strongly correlated to metabolic distances for Diet 1, 3 and Diet 4

The distributions of the order of colonization are then plotted for the ordered and random paths as seen in Figure 2 A, B, C, D. We see a good separation between the above networks for Diet 1, 3 and 4. This can be noticed as the distributions for the ordered network has lower jaccard's distance and is aligned more to the left when compared to the random network that has a greater jaccard's distance and is shifted more towards the right. An explanation to this can be that for Diet 1 (mother's milk) I see emergence of genera that are inter-related to each other in function based on the food source. However, if I shuffle the ordering of the organisms they might not be able to digest the byproducts of the food source in the order that they appear in the diet. Similarly, for Diet 3 and 4 that comprises of pre-adult and adult nutrition the organisms have to work in tandem to utilize the complex molecules generated in the intermittent steps of metabolic assimilation. On randomizing the genera and hence enzymes we tend

to break this co-operation pattern thereby increasing the metabolic distance between the organisms.

The fact to note is, that before Diet 2 (mother's milk and rice cereal) antibiotics was administered to the infant as the child had a bout of fever. This could hamper the abundance of certain organisms present in the baby and thus their enzymatic function. As a result, I see that the ordered path and the random path within the diet overlap. However, on calculating the Wilcoxon Rank test to determine whether the distributions are different I do see a significant P- value which states that the ordered path is indeed smaller in distance than the random path in the diet. This is demonstrated in Table 5.

Diet	Diet v/s Random	Diet v/s KEGG
Diet 1	1×10^{-101}	8×10^{-321}
Diet 2	1×10^{-14}	6×10^{-176}
Diet 3	8×10^{-93}	2×10^{-299}
Diet 4	1×10^{-94}	4×10^{-309}

Table 5: Calculated Wilcoxon P-Value between the ordered path v/s random path in a diet and ordered path and random path in KEGG

2.3.3 Organisms can differentiate between diets

Abundance of organisms was normalized across each time point by QIIME. This means that their total abundance adds up to 1 in a particular time step. Using this data and also overlapping the information of the days and diet I plotted a PCA plot. The aim was to see if the samples could be sorted out based on their diet labels Figure 3. From the

figure, it is evident that the organism content in each day can cluster well with time series data. Diet 1 and Diet 2 tend to cluster together and samples in Diet 3 and Diet 4 cluster near each other. There are some samples in Diet 3 that tend to group near Diet 1 and Diet 2. The anomaly could be explained by the fact that antibiotics was administered to the infant before collecting the samples. This could have upset the diversity of organisms in the beginning of that set of data.

On analyzing the alpha diversity of the samples, which measures the diversity of the organisms in a community/time step I also noticed that with a more complex diet the organism complexity and number of microbes increased Figure 4.

2.3.4 Enzymes can differentiate between diets

In order to show that network expansion approach using diet and KEGG enzyme information could produce any signal among different diets I also plotted the enzyme content in each time point labeled with the diet information. The enzyme information for the time points was obtained by looking at what organisms were present in a particular day and querying KEGG to see what they contributed enzymatically. This also showed the enzymes could indeed distinguish between diets Figure 5. This was not surprising as the organisms and the diets are closely related as shown in the section 2.3.3. This meant that network expansion approaches could be exploited to find differences between the diets.

2.3.5 Network Expansion results

Firstly, network expansion with different seed set cut-offs produced different sets of reactions to be turned on and off between diets shown in Supplementary Table 8 to 10. As seen in Supplementary Table 9 the log abundance of metabolites over 0 does not produce many reactions between diets. This proves that taking only abundant metabolites is not essential for a good network expansion. We need low abundant vitamins, molecules and co-factors like biotin for certain reactions to turn on [73]. Since the threshold of seed compounds over 0 produced few changes in the reaction among diets as shown in Supplementary table 9, I only concentrated in the network expansion results of the other 2 thresholds.

To demonstrate the success of the network expansion method with literature, I grouped the top 20 organisms that are closest to the network expansion profile given in Table 6.

Diet 1 organisms	Diet 2 organisms	Diet 3 organisms	Diet 4 organisms
Clostridium	Veillonella	Lactobacillus	Bacteroides
Streptococcus	Clostridium	Veillonella	Prevotella
Veillonella	Lactobacillus	Ruminococcus	Ruminococcus
Staphylococcus	Bifidobacterium	Clostridium	Faecalibacterium
Hafnia	Eubacterium	Enterococcus	Blautia
Enterococcus	Enterococcus	Prevotella	Coprococcus
Prevotella	Streptococcus	Streptococcus	Streptococcus
Corynebacterium	Blautia	Akkermansia	Clostridium
Gluconacetobacter	Enterobacter	Eubacterium	Roseburia
Bacteroides	Actinomyces	Faecalibacterium	Enterococcus
Bifidobacterium	Staphylococcus	Blautia	Veillonella
Enterobacter	Klebsiella	Bacteroides	Akkermansia
Klebsiella	Ruminococcus	Bifidobacterium	Eubacterium
Actinomyces	Rothia	Coprococcus	Parabacteroides
Rothia	Corynebacterium	Roseburia	Haemophilus
Finegoldia	Atopobium	Klebsiella	Pseudomonas
Campylobacter	Prevotella	Enterobacter	Megasphaera
Haemophilus	Coprococcus	Actinomyces	Alistipes
Anaerococcus	Citrobacter	Haemophilus	Hafnia
Roseburia	Eggerthella	Atopobium	Anaerostipes

Table 6: Top 20 organisms with threshold of -5 and no threshold

The thing to note is that these organisms are present in most samples of the dataset but they might necessarily not be the top candidates to resemble the expanded network. After I obtained a list of the top 20 organisms in each diet I looked at which organisms appeared to be unique to a diet based on the list. Those are marked in red in the table labelled as specialists in the diet. The ones in green were those organisms found across all the diets and was labelled as generalists. These were – *Streptococcus, Clostridium, Ruminococcus* (found in Diet 2 – 4), *Bifidobacterium* (found in Diet 1 – 3), *Veilonella, Prevotella* and *Enterococcus*. Literature suggests that the first 4 out of these are indeed found in most human gut microbiome [74].

I was more interested in the specialists in the diet and upon closer inspection of some organisms I found out that some genera in Diet 1 and Diet 4 were well correlated with certain foods. For Diet 1 among the 4 specialists found in the top 20, one of them was *Gluconacetobacter* that is a distinctly associated to the breast microbiome [75]. This means that this genus was transferred from the mother to the infant sample via the diet (milk) during breast-feeding. This also means that the algorithm can also tell us something about the environment of origin of the bacteria. Also found in the same diet was *Finegoldia* where studies have shown that this is mostly present in the infant during days 4-6 of the milk diet [76]. As a result, this one is not a top-ranking organism to the network expansion in Diet 1. Upon inspecting the adult diet (Diet 4) there were 5 specialists in the top 20 – *Parabacteroides, Megasphaera, Alistipes, Pseudomonas* and *Anaerostipes*. The first 3 out of those are strongly correlated to an animal diet with is why they show up in the adult diet and not in the pre-adult one [46 77].

Apart from the literature lookup I also performed a statistical analysis between the top 20 organisms in each diet with the aim to predict if these could differentiate between the diets. Using a Kolmogorov-Smirnov (KS) test and Euclidean distance of the microbe to a particular network expansion profile I was able to deduce that the distributions of organisms in Diet 1 were indeed different when compared to Diet 2 ($p\ value = 7.2\ x\ 10^{-4}$), Diet 3($p\ value = 2.4\ x\ 10^{-7}$) and Diet 4($p\ value = 5.5\ x\ 10^{-10}$) respectively as seen is Figure 6 and 7. This also shows that not only can the algorithm can distinguish between organisms in a diet but also the signal is more pronounced in pre-adult diet vs adult diet.

Figure 1: Phyla organization in the infant gut from QIIME analysis with legend

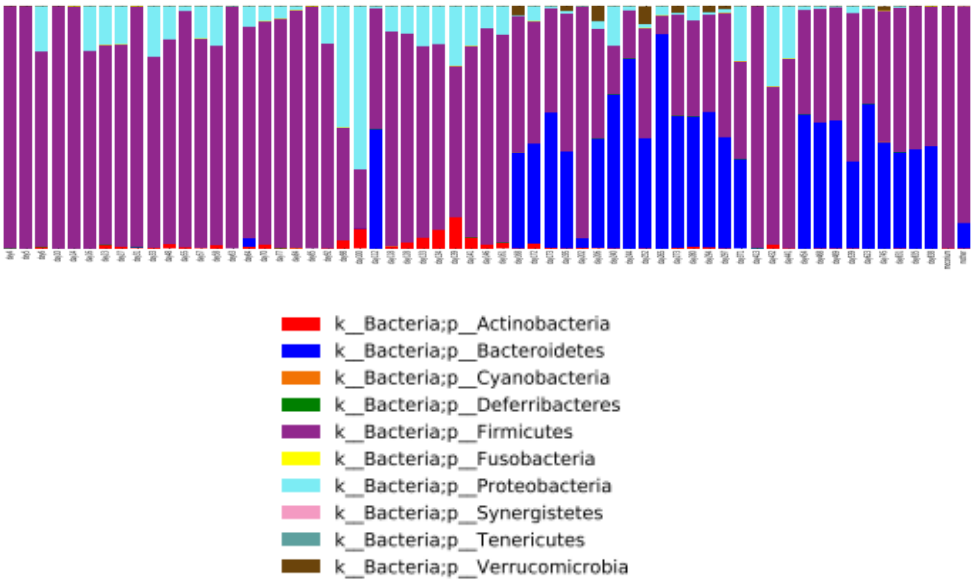


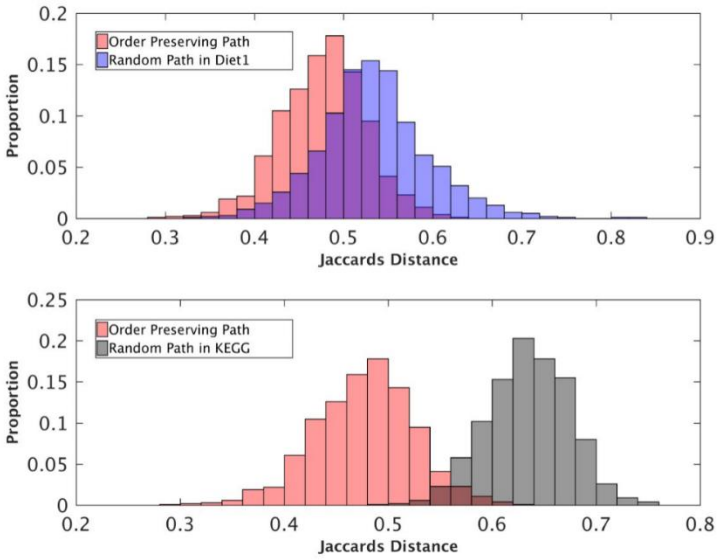
Figure 2: (A) Order of Colonization in ordered Diet 1 vs Random paths in Diet 1

(B) Order of Colonization in ordered Diet 2 vs Random paths in Diet 2

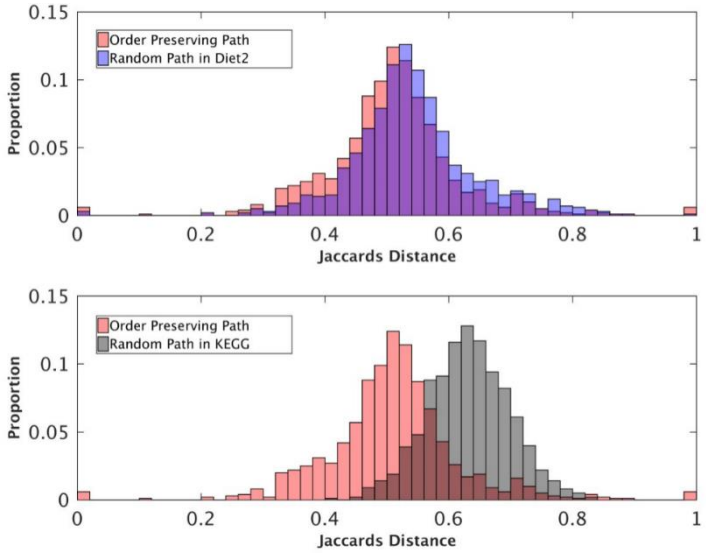
(C) Order of Colonization in ordered Diet 3 vs Random paths in Diet 3

(D) Order of Colonization in ordered Diet 4 vs Random paths in Diet 4

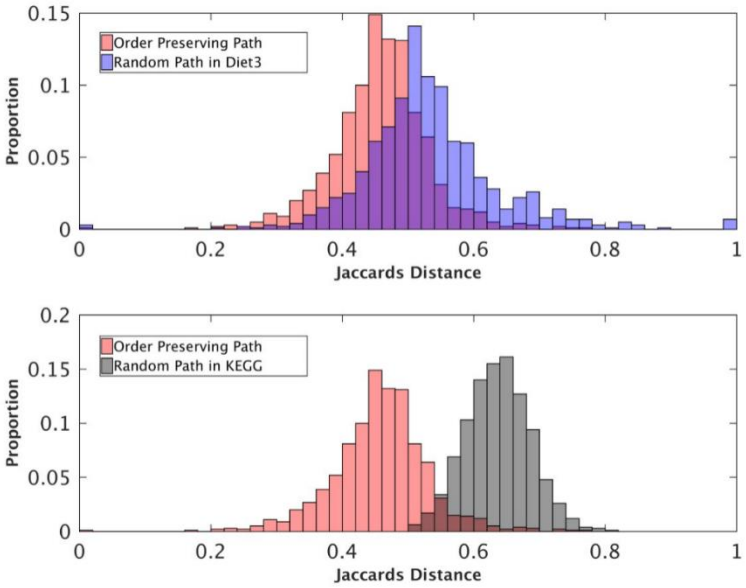
A



B

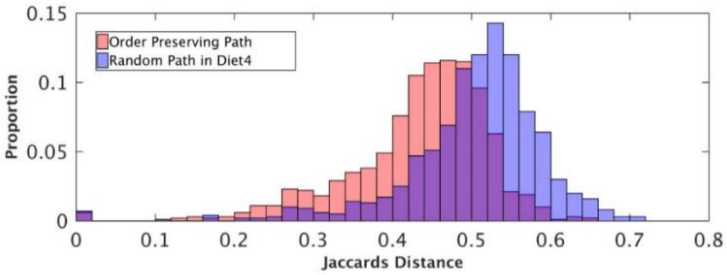


C


Order Preserving Path
Random Path in Diet3
Proportion
Jaccards Distance
0.15
0.1
0.05
0
0.2
0.4
0.6
0.8
1

Order Preserving Path
Random Path in KEGG
Proportion
Jaccards Distance
0.2
0.15
0.1
0.05
0
0.2
0.4
0.6
0.8
1

D


Order Preserving Path
Random Path in Diet4
Proportion
Jaccards Distance
0.15
0.1
0.05
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

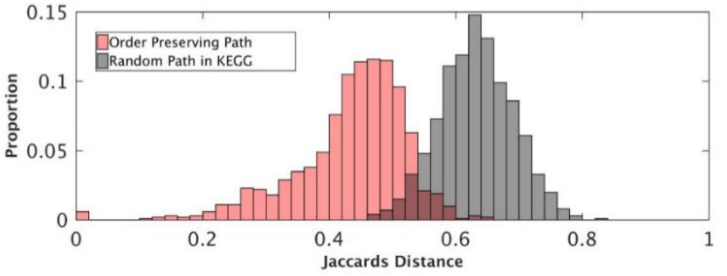

Order Preserving Path
Random Path in KEGG
Proportion
Jaccards Distance
0.15
0.1
0.05
0
0.2
0.4
0.6
0.8
1

Figure 3: OTUs can distinguish between diets

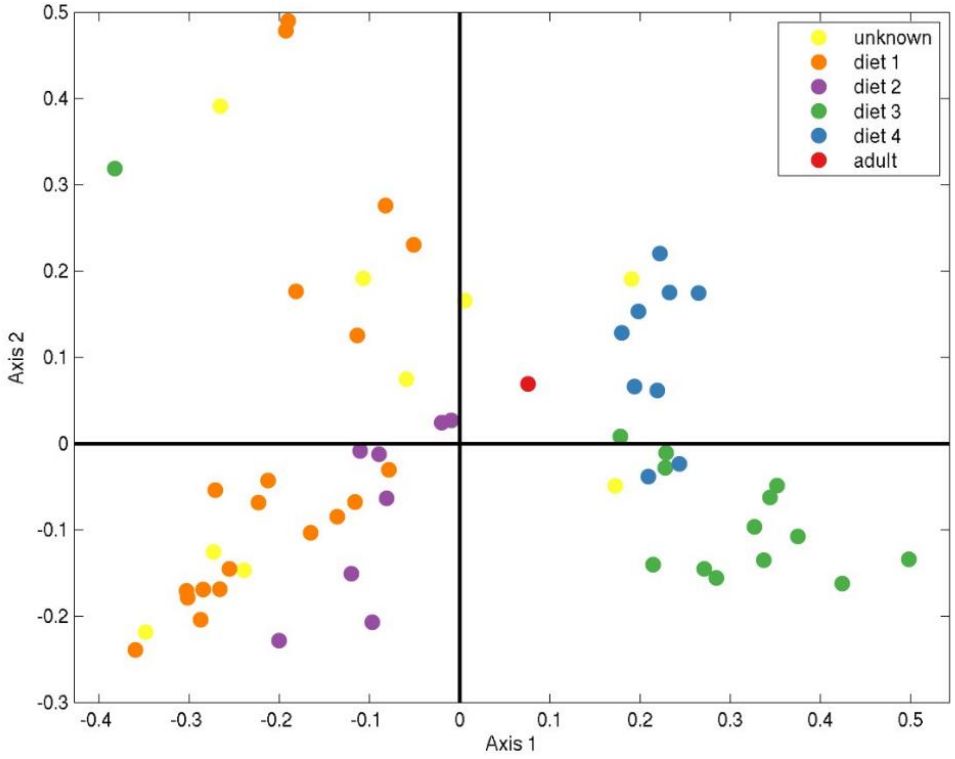


Figure 4: OTU diversity increases with the diet complexity

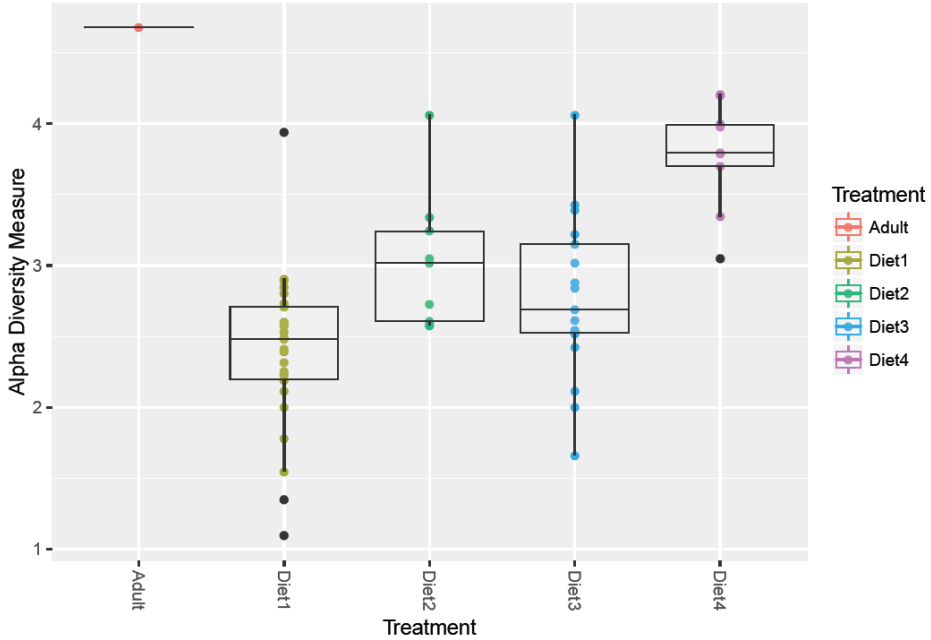


Figure 5: Enzymes can distinguish between diets

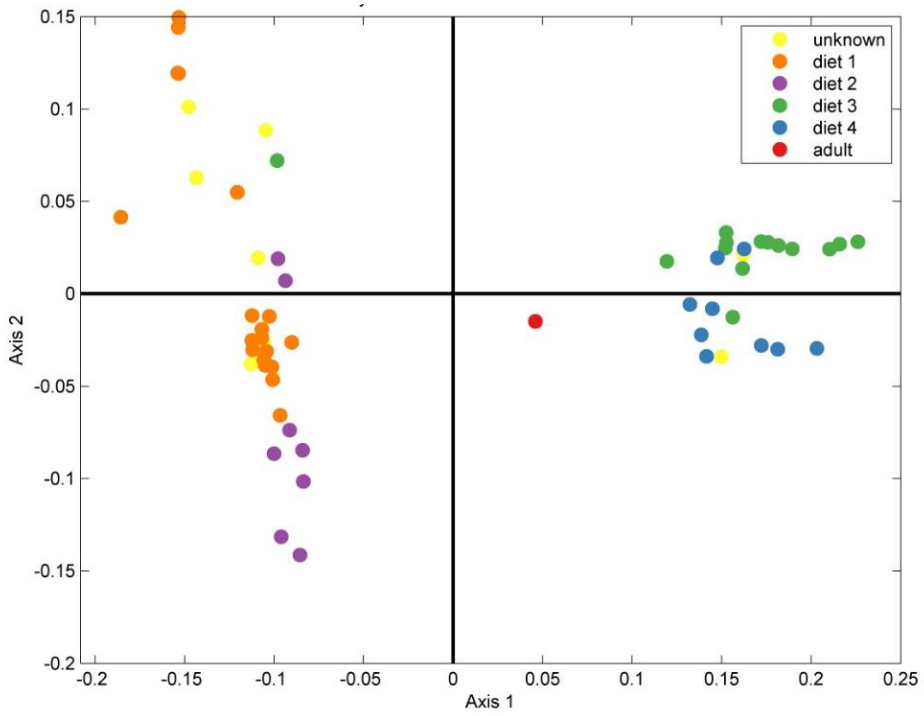


Figure 6: KS Test to show top 20 organisms can differentiate between diets – threshold -5

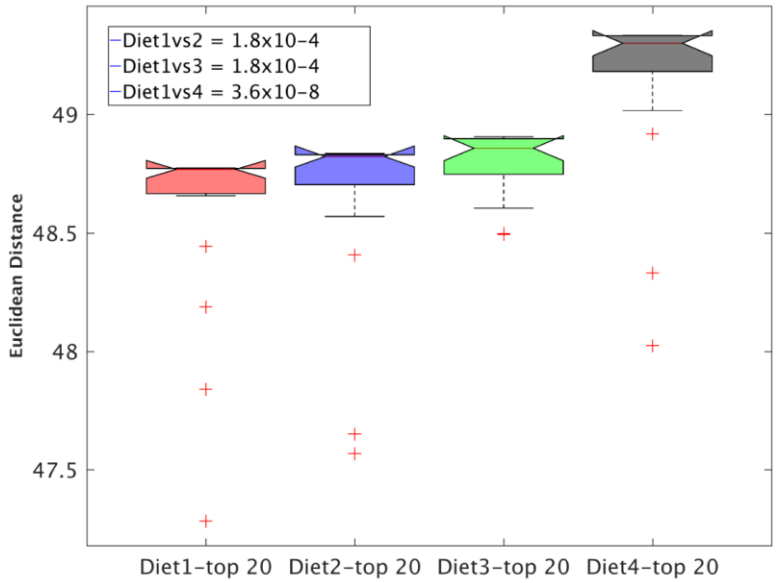
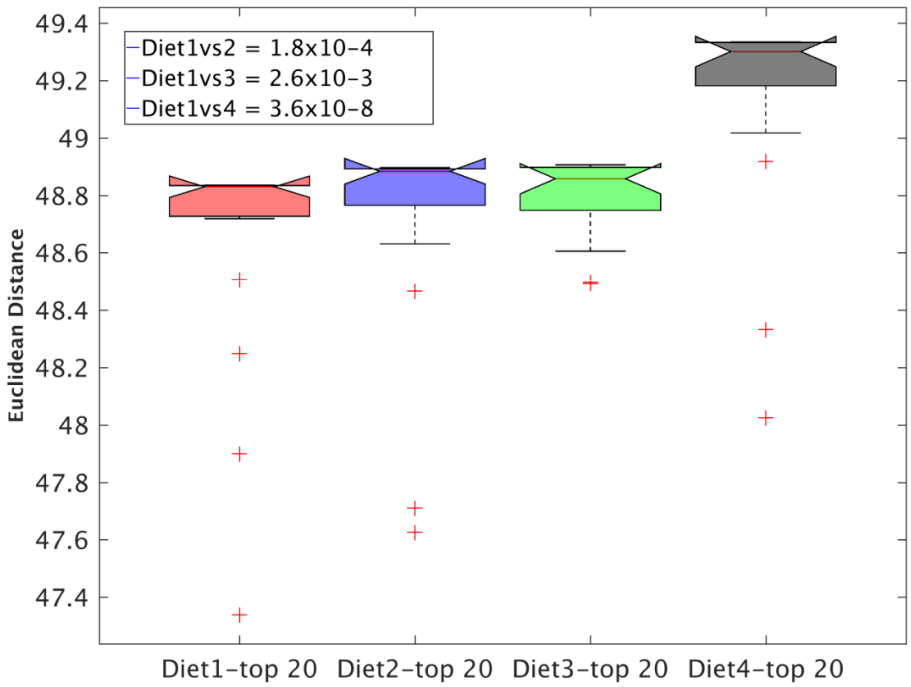


Figure 7: KS Test to show top 20 organisms can differentiate between diets – no threshold



2.4 Discussion

From all the above analysis, I can conclude that metabolic proximity does play an important role in determining the order of colonization. Since the random path always performs slightly worse in the enzymatic distance than the sequential or ordered path. Biologically this also makes sense as in the complex gut ecosystem the organisms exist mostly in cooperation and the early organisms prime the environment for the latter ones.

However, this analysis does not take into account multiple hosts. Though it does corroborate the signal found in the oral microbiome [42], further studies need to ascertain the strength of enzymes in the colonization pattern.

I can also see a clear trend in the abundance patterns of the organisms increase and decrease with certain food molecules suggesting that there is a correlation to the organism OTUs and its biochemical function. These OTUs and their enzymatic content can also distinguish between diets. This means that if diet curation is possible along with enzyme information of the organisms present, network expansion can be performed. This is a novel method to predict what microbes would resemble a diet closely. This is what I have done for the second part of this chapter and used literature to corroborate these findings.

Network expansion though partially successful has some issues that can be improved upon. The foremost thing is that literature reviews are only available for adult and neonatal diets. There isn't enough literature for specialized infant diets like Diet 2 and Diet 3. Also the curation of the diet was approximated to most infants specifically for Diets 2 – 4 including formula and amounts used. It was not specific for this dataset. The

other concern is the input compound and reaction seed set. Starting off with the compound set I chose to include only the low abundant metabolites. On using all the metabolites, the network for all diets saturate to the same result making this algorithm not very sensitive to the changes. However, a better approach of figuring out the importance of each compound would be to look at the biomass reaction of a genome scale flux balance model [78]. The biomass equation in such models provide information as to what molecules are important for the organism growth. This can be used to set the threshold for the compounds in the seed set. In general, there are 2 types of biomass reactions – gram positive and gram negative microbes which require different compounds to survive and different thresholds for these. Also, there are several assumptions made for the reaction set. The co-factors were removed from the stoichiometric matrix and there are many special organism reactions which were not considered. These modifications could increase the signal and sensitivity of this algorithm.

Supplementary

Table 1: Complete metabolite set of Diet 1 (Mother's Milk)

Compounds	Names	Total
C00001	Water	700.7
C00023	Iron	0.00026
C00034	Manganese	0.000208
C00038	Zinc	0.0013
C00070	Copper	0.000416
C00072	Vitamin C	0.039
C00076	Calcium	0.26
C00114	Choline	0.1274
C00120	Biotin	6.57E-06
C00187	Cholesterol	0.104
C00219	Arachidonic acid	0.208
C00238	Potassium	0.416
C00243	Lactose	55.12
C00249	Palmitic acid	7.358
C00253	Niacin	0.00143
C00255	Riboflavin	0.000286
C02679	Lauric acid	2.054
C00305	Magnesium	0.026
C00314	Pyridoxine	0.000078
C00378	Thiamin(e)	0.000104
C00473	Vitamin A	0.001001
C00504	Folate	0.000052
C00698	Chloride	1.001
C00712	Oleic acid	11.804

C00719	Betaine	0.09204
C00864	Pantothenic acid	0.001794
C01330	Sodium	0.13
C01382	Iodine	4.77E-05
C01529	Selenium	1.82E-05
C01530	Stearic acid	2.34
C01571	Capric acid	0.494
C01595	Linoleic acid	2.99
C01628	Vitamin K	2.6E-06
C05441	Vitamin D2	5.2E-06
C05443	Vitamin D3	7.8E-06
C05776	Vitamin B12	5.2E-07
C06262	Phosphorus	0.104
C06424	Myristic acid	2.574
C06427	Linolenic acid	0.416
C08362	Palmitoleic acid	1.04
C16526	Eicosenoic acid	0.312
C00078	Tryptophan	0.13
C00188	Threonine	0.364
C00407	Isoleucine	0.442
C00123	Leucine	0.754
C00047	Lysine	0.546
C00073	Methionine	0.156
C00097	Cysteine	0.156
C00079	Phenylalanine	0.364
C00082	Tyrosine	0.416
C00183	Valine	0.494

C00062	Arginine	0.338
C00135	Histidine	0.182
C00041	Alanine	0.286
C00049	Aspartic acid	0.65
C00025	Glutamic acid	1.352
C00037	Glycine	0.208
C00148	Proline	0.65
C00065	Serine	0.338

Table 2: Complete metabolite set of Diet 2 (Mother's Milk + Rice Cereal)

Compounds	Names	Total
C00041	Alanine	0.308
C00219	Arachidonic acid	0.224
C00062	Arginine	0.364
C00049	Aspartic acid	0.7
C00719	Betaine	0.09912
C00120	Biotin	7.08E-06
C00076	Calcium	0.291
C01571	Capric acid	0.532
C00698	Chloride	1.078
C00187	Cholesterol	0.112
C00114	Choline	0.1374
C00070	Copper	0.000451
C00097	Cysteine	0.168
C16526	Eicosenoic acid	0.336
C00504	Folate	0.000059

C00031	Glucose	0.03
C00025	Glutamic acid	1.456
C00037	Glycine	0.224
C00135	Histidine	0.196
C01382	Iodine	5.13E-05
C00023	Iron	0.000945
C00407	Isoleucine	0.476
C00243	Lactose	59.36
C02679	Lauric acid	2.212
C00123	Leucine	0.812
C01595	Linoleic acid	3.229
C06427	Linolenic acid	0.449
C00047	Lysine	0.588
C00305	Magnesium	0.0285
C00208	Maltose	0.025
C00034	Manganese	0.000242
C00073	Methionine	0.168
C06424	Myristic acid	2.772
C00253	Niacin	0.001828
C00712	Oleic acid	12.717
C00249	Palmitic acid	7.926
C08362	Palmitoleic acid	1.12
C00864	Pantothenic acid	0.001952
C00079	Phenylalanine	0.392
C06262	Phosphorus	0.1155
C00238	Potassium	0.4515
C00148	Proline	0.7

C00314	Pyridoxine	0.000095
C00255	Riboflavin	0.000326
C01529	Selenium	1.97E-05
C00065	Serine	0.364
C01330	Sodium	0.1405
C01530	Stearic acid	2.5205
C00089	Sucrose	0.005
C00378	Thiamin(e)	0.000127
C00188	Threonine	0.392
C00078	Tryptophan	0.14
C00082	Tyrosine	0.448
C00183	Valine	0.532
C00473	Vitamin A	0.001078
C05776	Vitamin B12	6.2E-07
C00072	Vitamin C	0.04205
C05441	Vitamin D2	5.6E-06
C05443	Vitamin D3	8.5E-06
C02477	Vitamin E	0.00006
C01628	Vitamin K	2.8E-06
C00001	Water	754.665
C00038	Zinc	0.001535
C00317	Amylopectin	1.0286
C00718	Amylose	0.3614

Table 3: Complete metabolite set of Diet 3 (Mother's Milk + Rice Cereal + Formula + Peas)

Compounds	Names	Total
C00001	Water	1760.985
C00023	Iron	0.004125
C00025	Glutamic acid	1.502881
C00031	Glucose	0.245968
C00034	Manganese	0.00018
C00037	Glycine	0.224128
C00038	Zinc	0.002542
C00041	Alanine	0.300678
C00047	Lysine	0.573369
C00049	Aspartic Acid	0.625048
C00062	Arginine	0.336535
C00065	Serine	0.356636
C00070	Copper	0.000474
C00072	Vitamin C	0.044402
C00073	Methionine	0.169072
C00076	Calcium	0.33692
C00078	Tryptophan	0.142027
C00079	Phenylalanine	0.408421
C00082	Tyrosine	0.460421
C00089	Sucrose	1.793331
C00095	Fructose	0.138206
C00097	Cysteine	0.181072
C00718	Amylose	2.4219
C00114	Choline	0.13624
C00120	Biotin	2.99E-05

C00123	Leucine	0.778323
C00124	Galactose	0.00179
C00135	Histidine	0.180276
C00137	Inositol	0.000459
C00148	Proline	0.663926
C00182	Valine	0.483794
C00187	Cholesterol	0.096
C00188	Threonine	0.383498
C00208	Maltose	0.185244
C00219	Arachidonic acid	0.20344
C00238	Potassium	0.48974
C00243	Lactose	55.17409
C00249	Palmitic acid	7.31664
C00253	Niacin	0.003622
C00255	Riboflavin	0.000443
C02679	Lauric acid	2.07816
C00305	Magnesium	0.04002
C00314	Pyridoxine	0.000185
C00317	Amylopectin	6.8931
C00378	Thiamin(e)	0.000259
C00407	Isoleucine	0.447854
C00473	Vitamin A	0.000967
C00504	Folate	8.92E-05
C00698	Chloride	0.948499
C00712	Oleic acid	11.709
C00719	Betaine	0.085066
C00742	Fluoride	1.02E-05

C00864	Pantothenic acid	0.002151
C01330	Sodium	0.16106
C01382	Iodine	4.40E-05
C01529	Selenium	1.82E-05
C01530	Stearic acid	2.25498
C01571	Capric acid	0.47888
C01585	Caporic acid	0.00528
C01595	Linoleic acid	3.304208
C01628	Vitamin K	0.046006
C02477	Vitamin E	0.001024
C02483	gamma-Tocopherol	0.000337
C05441	Vitamin D2	5.11E-06
C05443	Vitamin D3	8.01E-06
C05776	Vitamin B12	8.94E-07
C06262	Phosphorus	0.16822
C06423	Caprylic acid	0.03344
C06424	Myristic acid	2.46048
C06427	Linolenic acid	0.46244
C08601	Lutein	0.001024
C08316	Erucic acid	0.00528
C08362	Palmitoleic acid	0.96528
C16526	Eicosenoic acid	0.29328

Table 4: Complete metabolite set of Diet 3 (Cow's Milk + Adult Diet)

Compounds	Names	Total
C09727	(-)-Epicatechin	0.0005
C12136	(-)-Epigallocatechin	0.00087
C06562	(+)-Catechin	0.102
C00041	Alanine	1.237356
C00317	Amylopectin	24.6639
C00718	Amylose	23.902
C01889	Arabinoxylan	0.1719
C06425	Arachidic acid	0.0402
C00219	Arachidonic acid	0.0598
C00062	Arginine	1.341628
C00049	Aspartic acid	2.726346
C00965	Beta glucans	0.34825
C01753	beta sitosterol	0.008206
C00719	Betaine	0.0238
C01753	Beta-sitosterol	0.001
C00120	Biotin	2.29E-05
C00246	Butyric acid	0.6716
C00076	Calcium, Ca	0.319
C01789	Campesterol	0.001919
C01571	Capric acid	0.392
C01585	Caproic acid	0.4042
C06423	Caprylic acid	0.185
C05433	Carotene, alpha	0.000001
C02094	Carotene, beta	1.18E-04
C00760	Cellulose	1.6541

C00698	Chloride	0.21056
C00187	Cholesterol	0.17
C00114	Choline	0.18411
C16513	clupanodonic acid	0.001
C05776	Cobalmin(III)	2.00E-08
C00070	Copper, Cu	0.000927
C08591	Cryptoxanthin, beta	1.13E-05
C05905	Cyanidin	0.00087
C00491	Cystine	0.428285
C01571	Decanoic	0.3778
C06429	Docosahexaenoic acid (DHA)	0.011
C08281	docosanoic acid	0.117
C16513	Docosapentaenoic acid	0.003
C08316	Docosenoic acid	0.984
C16525	Eicosadienoic acid	0.003
C06428	Eicosapentaenoic acid (EPA)	0.003
C16522	Eicosatrienoic acid	0.006
C16526	Eicosenoic acid	0.987
C00742	Fluoride, F	5.51E-05
C00504	Folate, total	2.32E-04
C01355	Fructan	0.01365
C00095	Fructose	6.414
C00124	Galactose	0.8585
C00333	Galacturnic acid	0.0074
C06426	gamma-linolenic acid	0.004
C06563	Genistein	0.01
C00031	Glucose (dextrose)	6.838

C00025	Glutamic acid	4.046505
C00037	Glycine	1.107421
C15805	Guaiacyl lignin	0.0255
C16536	Heptadecanoic acid	0.1444
C00135	Histidine	0.593237
C01157	Hydroxyproline	0.001
C01382	Iodine	7.33E-05
C00023	Iron, Fe	0.006218
C08821	Isofucosterol	0.000505
C00407	Isoleucine	1.098167
C00243	Lactose	3.546
C02679	Lauric acid	0.6624
C00123	Leucine	1.927271
C08320	Lignoceric acid	0.001
C01595	Linoleic acid	4.7288
C06427	Linolenic (alpha) acid	0.741
C08601	Lutein + zeaxanthin	1.36E-04
C00047	Lysine	1.572039
C00305	Magnesium, Mg	0.089
C00208	Maltose	0.32
C00034	Manganese, Mn	0.016171
C00159	Mannose	0.1015
C00073	Methionine	0.425867
C06424	Myristic acid	1.8354
C06424	Myristoleic acid	0.0662
C08323	nervonic acid	0.001
C00253	Niacin	0.008448

C01530	Octadecadienoic acid	0.3872
C06427	Octadecatrienoic acid	0.007
C00712	Oleic acid	6.77
C00249	Palmitic acid	7.0016
C08362	Palmitoleic acid	0.8916
C00864	Pantothenic acid	0.003907
C00714	Pectin	0.0648
C16537	Pentadecanoic	0.0976
C00079	Phenylalanine	1.118356
C06262	Phosphorus, P	0.5092
C15804	p-hydroxyphenyl lignin	0.0255
C00238	Potassium, K	0.6556
C10237	Proanthocyanidin A2 (dimer)	0.3006
C00148	Proline	1.440016
C00389	Quercetin	0.00087
C00492	Raffinose	0.12356
C00507	Rhammose	0.216
C00255	Riboflavin	0.002943
C01529	Selenium, Se	6.86E-05
C00065	Serine	1.050668
C01330	Sodium, Na	0.71555
C00794	Sorbitol	0.57
C01613	Stachyose	0.3222
C01530	Stearic acid	2.771
C05442	Stigmasterol	0.001641
C00089	Sucrose	9.5432
C15806	Syringyl lignin	0.0255

C08320	Tetracosanoic acid	0.019
C00378	Thiamin	0.00252
C00188	Threonine	0.996535
C06428	timnodonic acid	0.014
C14152	Tocopherol, beta	0.00007
C14151	Tocopherol, delta	0.00028
C02483	Tocopherol, gamma	0.00301
C00078	Tryptophan	0.316556
C00082	Tyrosine	0.898684
C00182	Valine	1.295393
C08252	Verbascose	0.37
C00473	Vitamin A, RAE	0.001814
C05776	Vitamin B-12	34
C00314	Vitamin B-6 (pyridoxine)	0.002621
C00072	Vitamin C	0.00503
C01673	Vitamin D	2.73E-05
C05441	Vitamin D2	2.5E-07
C05433	Vitamin D3	1.98E-06
C02477	Vitamin E (alpha-tocopherol)	0.003474
C01628	Vitamin K (phylloquinone)	1.00E-05
C00001	Water	261.962
C00181	Xylose	0.324
C00038	Zinc, Zn	0.018892

Table 5: Breakup of complex molecules – Casein

C00078	Tryptophan
C00188	Threonine
C00407	Isoleucine
C00123	Leucine
C00047	Lysine
C00073	Methionine
C00097	Cysteine
C00079	Phenylalanine
C00082	Tyrosine
C00183	Valine
C00062	Arginine
C00135	Histidine
C00041	Alanine
C00049	Aspartic acid
C00025	Glutamic acid
C00037	Glycine
C00148	Proline
C00065	Serine

Table 6: Seed set for metabolite abundance cut-off above -5

Diet 1		**Diet 2**		**Diet 3**		**Diet 4**	
C00001	Water	C00041	Alanine	C00001	Water	C09727	(-)-Epicatechin
C00023	Iron	C00219	Arachidonic acid	C00023	Iron	C12136	(-)-Epigallocatechin
C00034	Manganese	C00062	Arginine	C00025	Glutamic acid	C06562	(+)-Catechin
C00038	Zinc	C00049	Aspartic acid	C00031	Glucose	C00041	Alanine
C00070	Copper	C00719	Betaine	C00034	Manganese	C00317	Amylopectin
C00072	Vitamin C	C00076	Calcium	C00037	Glycine	C00718	Amylose
C00076	Calcium	C01571	Capric acid	C00038	Zinc	C01889	Arabinoxylan
C00114	Choline	C00698	Chloride	C00041	Alanine	C06425	Arachidic acid

C00187	Cholesterol	C00187	Cholesterol	C00047	Lysine	C00219	Arachidonic acid
C00219	Arachidonic acid	C00114	Choline	C00049	Aspartic Acid	C00062	Arginine
C00238	Potassium	C00070	Copper	C00062	Arginine	C00049	Aspartic acid
C00243	Lactose	C00097	Cysteine	C00065	Serine	C00965	Beta glucans
C00249	Palmitic acid	C16526	Eicosenoic acid	C00070	Copper	C01753	beta sitosterol
C00253	Niacin	C00504	Folate	C00072	Vitamin C	C00719	Betaine
C00255	Riboflavin	C00031	Glucose	C00073	Methionine	C01753	Beta-sitosterol
C02679	Lauric acid	C00025	Glutamic acid	C00076	Calcium	C00120	Biotin
C00305	Magnesium	C00037	Glycine	C00078	Tryptophan	C00246	Butyric acid
C00314	Pyridoxine	C00135	Histidine	C00079	Phenylalanine	C00076	Calcium, Ca
C00378	Thiamin(e)	C01382	Iodine	C00082	Tyrosine	C01789	Campesterol
C00473	Vitamin A	C00023	Iron	C00089	Sucrose	C01571	Capric acid
C00504	Folate	C00407	Isoleucine	C00095	Fructose	C01585	Caproic acid
C00698	Chloride	C00243	Lactose	C00097	Cysteine	C06423	Caprylic acid
C00712	Oleic acid	C02679	Lauric acid	C00718	Amylose	C02094	Carotene, beta
C00719	Betaine	C00123	Leucine	C00114	Choline	C00760	Cellulose
C00864	Pantothenic acid	C01595	Linoleic acid	C00120	Biotin	C00698	Chloride
C01330	Sodium	C06427	Linolenic acid	C00123	Leucine	C00187	Cholesterol
C01382	Iodine	C00047	Lysine	C00124	Galactose	C00114	Choline
C01529	Selenium	C00305	Magnesium	C00135	Histidine	C16513	clupanodonic acid
C01530	Stearic acid	C00208	Maltose	C00137	Inositol	C00070	Copper, Cu
C01571	Capric acid	C00034	Manganese	C00148	Proline	C08591	Cryptoxanthin, beta
C01595	Linoleic acid	C00073	Methionine	C00182	Valine	C05905	Cyanidin
C06262	Phosphorus	C06424	Myristic acid	C00187	Cholesterol	C00491	Cystine
C06424	Myristic acid	C00253	Niacin	C00188	Threonine	C01571	Decanoic
C06427	Linolenic acid	C00712	Oleic acid	C00208	Maltose	C06429	Docosahexaenoic acid (DHA)

C08362	Palmitoleic acid	C00249	Palmitic acid	C00219	Arachidonic acid	C08281	docosanoic acid
C16526	Eicosenoic acid	C08362	Palmitoleic acid	C00238	Potassium	C16513	Docosapentaenoic acid
C00078	Tryptophan	C00864	Pantothenic acid	C00243	Lactose	C08316	Docosenoic acid
C00188	Threonine	C00079	Phenylalanine	C00249	Palmitic acid	C16525	Eicosadienoic acid
C00407	Isoleucine	C06262	Phosphorus	C00253	Niacin	C06428	Eicosapentaenoic acid (EPA)
C00123	Leucine	C00238	Potassium	C00255	Riboflavin	C16522	Eicosatrienoic acid
C00047	Lysine	C00148	Proline	C02679	Lauric acid	C16526	Eicosenoic acid
C00073	Methionine	C00314	Pyridoxine	C00305	Magnesium	C00742	Fluoride, F
C00097	Cysteine	C00255	Riboflavin	C00314	Pyridoxine	C00504	Folate, total
C00079	Phenylalanine	C01529	Selenium	C00317	Amylopectin	C01355	Fructan
C00082	Tyrosine	C00065	Serine	C00378	Thiamin(e)	C00095	Fructose
C00183	Valine	C01330	Sodium	C00407	Isoleucine	C00124	Galactose
C00062	Arginine	C01530	Stearic acid	C00473	Vitamin A	C00333	Galacturnic acid
C00135	Histidine	C00089	Sucrose	C00504	Folate	C06426	gamma-linolenic acid
C00041	Alanine	C00378	Thiamin(e)	C00698	Chloride	C06563	Genistein
C00049	Aspartic acid	C00188	Threonine	C00712	Oleic acid	C00031	Glucose (dextrose)
C00025	Glutamic acid	C00078	Tryptophan	C00719	Betaine	C00025	Glutamic acid
C00037	Glycine	C00082	Tyrosine	C00742	Fluoride	C00037	Glycine
C00148	Proline	C00183	Valine	C00864	Pantothenic acid	C15805	Guaiacyl lignin
C00065	Serine	C00473	Vitamin A	C01330	Sodium	C16536	Heptadecanoic acid
-	-	C00072	Vitamin C	C01382	Iodine	C00135	Histidine
-	-	C02477	Vitamin E	C01529	Selenium	C01157	Hydroxyproline
-	-	C00001	Water	C01530	Stearic acid	C01382	Iodine
-	-	C00038	Zinc	C01571	Capric acid	C00023	Iron, Fe

-	-	C00317	Amylopectin	C01585	Caporic acid	C08821	Isofucosterol
-	-	C00718	Amylose	C01595	Linoleic acid	C00407	Isoleucine
-	-	-	-	C01628	Vitamin K	C00243	Lactose
-	-	-	-	C02477	Vitamin E	C02679	Lauric acid
-	-	-	-	C02483	gamma-Tocopherol	C00123	Leucine
-	-	-	-	C06262	Phosphorus	C08320	Lignoceric acid
-	-	-	-	C06423	Caprylic acid	C01595	Linoleic acid
-	-	-	-	C06424	Myristic acid	C06427	Linolenic (alpha) acid
-	-	-	-	C06427	Linolenic acid	C08601	Lutein + zeaxanthin
-	-	-	-	C08601	Lutein	C00047	Lysine
-	-	-	-	C08316	Erucic acid	C00305	Magnesium, Mg
-	-	-	-	C08362	Palmitoleic acid	C00208	Maltose
-	-	-	-	C16526	Eicosenoic acid	C00034	Manganese, Mn
-	-	-	-	-	-	C00159	Mannose
-	-	-	-	-	-	C00073	Methionine
-	-	-	-	-	-	C06424	Myristic acid
-	-	-	-	-	-	C06424	Myristoleic acid
-	-	-	-	-	-	C08323	nervonic acid
-	-	-	-	-	-	C00253	Niacin
-	-	-	-	-	-	C01530	Octadecadienoic acid
-	-	-	-	-	-	C06427	Octadecatrienoic acid
-	-	-	-	-	-	C00712	Oleic acid
-	-	-	-	-	-	C00249	Palmitic acid
-	-	-	-	-	-	C08362	Palmitoleic acid
-	-	-	-	-	-	C00864	Pantothenic acid

-	-	-	-	-	-	C00714	Pectin
-	-	-	-	-	-	C16537	Pentadecanoic
-	-	-	-	-	-	C00079	Phenylalanine
-	-	-	-	-	-	C06262	Phosphorus, P
-	-	-	-	-	-	C15804	p-hydroxyphenyl lignin
-	-	-	-	-	-	C00238	Potassium, K
-	-	-	-	-	-	C10237	Proanthocyanidin A2 (dimer)
-	-	-	-	-	-	C00148	Proline
-	-	-	-	-	-	C00389	Quercetin
-	-	-	-	-	-	C00492	Raffinose
-	-	-	-	-	-	C00507	Rhammose
-	-	-	-	-	-	C00255	Riboflavin
-	-	-	-	-	-	C01529	Selenium, Se
-	-	-	-	-	-	C00065	Serine
-	-	-	-	-	-	C01330	Sodium, Na
-	-	-	-	-	-	C00794	Sorbitol
-	-	-	-	-	-	C01613	Stachyose
-	-	-	-	-	-	C01530	Stearic acid
-	-	-	-	-	-	C05442	Stigmasterol
-	-	-	-	-	-	C00089	Sucrose
-	-	-	-	-	-	C15806	Syringyl lignin
-	-	-	-	-	-	C08320	Tetracosanoic acid
-	-	-	-	-	-	C00378	Thiamin
-	-	-	-	-	-	C00188	Threonine
-	-	-	-	-	-	C06428	timnodonic acid
-	-	-	-	-	-	C14152	Tocopherol, beta
-	-	-	-	-	-	C14151	Tocopherol, delta

-	-	-	-	-	-	C02483	Tocopherol, gamma
-	-	-	-	-	-	C00078	Tryptophan
-	-	-	-	-	-	C00082	Tyrosine
-	-	-	-	-	-	C00182	Valine
-	-	-	-	-	-	C08252	Verbascose
-	-	-	-	-	-	C00473	Vitamin A, RAE
-	-	-	-	-	-	C05776	Vitamin B-12
-	-	-	-	-	-	C00314	Vitamin B-6 (pyridoxine)
-	-	-	-	-	-	C00072	Vitamin C
-	-	-	-	-	-	C01673	Vitamin D
-	-	-	-	-	-	C02477	Vitamin E (alpha-tocopherol)
-	-	-	-	-	-	C00001	Water
-	-	-	-	-	-	C00181	Xylose
-	-	-	-	-	-	C00038	Zinc, Zn

Table 7: Seed set for metabolite abundance cut-off above 0

Diet 1		**Diet 2**		**Diet 3**		**Diet 4**	
C00001	Water	C00698	Chloride	C00001	Water	C00041	Alanine
C00243	Lactose	C00025	Glutamic acid	C00025	Glutamic acid	C00317	Amylopectin
C00249	Palmitic acid	C00243	Lactose	C00089	Sucrose	C00718	Amylose
C02679	Lauric acid	C02679	Lauric acid	C00718	Amylose	C00062	Arginine
C00698	Chloride	C01595	Linoleic acid	C00243	Lactose	C00049	Aspartic acid
C00712	Oleic acid	C06424	Myristic acid	C00249	Palmitic acid	C00760	Cellulose
C01530	Stearic acid	C00712	Oleic acid	C02679	Lauric acid	C00095	Fructose
C01595	Linoleic acid	C00249	Palmitic acid	C00317	Amylopectin	C00031	Glucose (dextrose)

C06424	Myristic acid	C08362	Palmitoleic acid	C00712	Oleic acid	C00025	Glutamic acid
C08362	Palmitoleic acid	C01530	Stearic acid	C01530	Stearic acid	C00037	Glycine
C00025	Glutamic acid	C00001	Water	C01595	Linoleic acid	C00407	Isoleucine
-	-	C00317	Amylopectin	C06424	Myristic acid	C00243	Lactose
-	-	-	-	-	-	C00123	Leucine
-	-	-	-	-	-	C01595	Linoleic acid
-	-	-	-	-	-	C00047	Lysine
-	-	-	-	-	-	C06424	Myristic acid
-	-	-	-	-	-	C00712	Oleic acid
-	-	-	-	-	-	C00249	Palmitic acid
-	-	-	-	-	-	C00079	Phenylalanine
-	-	-	-	-	-	C00148	Proline
-	-	-	-	-	-	C00065	Serine
-	-	-	-	-	-	C01530	Stearic acid
-	-	-	-	-	-	C00089	Sucrose
-	-	-	-	-	-	C00182	Valine
-	-	-	-	-	-	C05776	Vitamin B-12
-	-	-	-	-	-	C00001	Water

Figure 1: Network Size v/s the cut-offs. The orange lines indicate the points used for case study

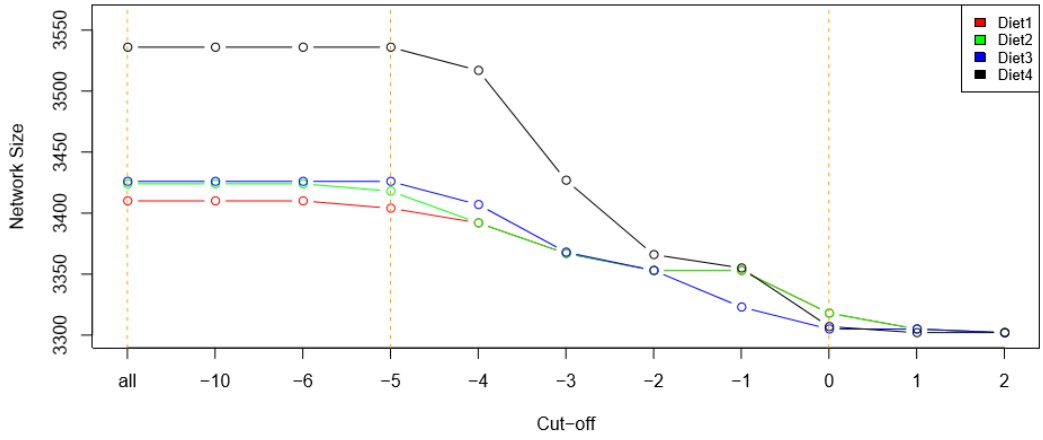


Table 8: Reactions added to the network with different diets for log abundance threshold of metabolites above -5

Cut-off	diet 1 - diet 2	diet 2 - diet 3	diet 3 - diet 4
-5	2.1.1.95	2.5.1.63	1.1.1.219
	2.1.1.295	3.5.1.12	1.3.1.77
	2.5.1.115	6.2.1.11	1.3.99.27
	2.5.1.116	6.3.4.9	1.17.1.3
	2.5.1.117	6.3.4.10	2.1.7.6
	5.5.1.24	6.3.4.11	2.1.1.82
	-	6.3.4.15	2.1.1.83
	-	-	2.1.1.210
	-	-	2.3.1.116
	-	-	2.3.1.153
	-	-	2.3.1.171
	-	-	2.3.1.172
	-	-	2.3.1.215

	-	-	2.4.1.91
	-	-	2.4.1.115
	-	-	2.4.1.116
	-	-	2.4.1.159
	-	-	2.4.1.237
	-	-	2.4.1.238
	-	-	2.4.1.239
	-	-	2.4.1.240
	-	-	2.4.1.249
	-	-	2.4.1.254
	-	-	2.4.1.294
	-	-	2.4.1.295
	-	-	2.4.1.297
	-	-	2.4.1.298
	-	-	2.4.2.35
	-	-	2.4.2.50
	-	-	2.4.2.51
	-	-	2.8.2.25
	-	-	2.8.2.26
	-	-	2.8.2.27
	-	-	2.8.2.28
	-	-	3.2.1.66
	-	-	4.2.1.131
	-	-	5.2.1.5
	-	-	5.2.1.13
	-	-	5.2.1.14
	-	-	5.3.3.13
	-	-	5.5.1.18
	-	-	5.5.1.19

Table 9: Reactions added to the network with different diets for log abundance threshold of metabolites above 0

Cut-off	diet 1 - diet 2	diet 2 - diet 3	diet 3 - diet 4
0	-	1.11.2.2	5.2.1.5
	-	2.5.1.94	-
	-	3.8.1.16	-
	-	3.8.1.17	-
	-	4.5.1.4	-
	-	5.2.1.10	-
	-	6.2.1.33	-

Table 10: Reactions added to the network with different diets for all metabolites

Cut-off	diet 1 - diet 2	diet 2 - diet 3	diet 3 - diet 4
all	2.1.1.95	2.5.1.63	1.1.1.219
	2.1.1.295	-	1.3.1.77
	2.5.1.115	-	1.3.99.27
	2.5.1.116	-	1.17.1.3
	2.5.1.117	-	2.1.7.6
	5.5.1.24	-	2.1.1.82
	-	-	2.1.1.83
	-	-	2.1.1.210
	-	-	2.3.1.116
	-	-	2.3.1.153
	-	-	2.3.1.171
	-	-	2.3.1.172
	-	-	2.3.1.215
	-	-	2.4.1.91
	-	-	2.4.1.115
	-	-	2.4.1.116
	-	-	2.4.1.159
	-	-	2.4.1.237

	-	-	2.4.1.238
	-	-	2.4.1.239
	-	-	2.4.1.240
	-	-	2.4.1.249
	-	-	2.4.1.254
	-	-	2.4.1.294
	-	-	2.4.1.295
	-	-	2.4.1.297
	-	-	2.4.1.298
	-	-	2.4.2.35
	-	-	2.4.2.50
	-	-	2.4.2.51
	-	-	2.8.2.25
	-	-	2.8.2.26
	-	-	2.8.2.27
	-	-	2.8.2.28
	-	-	3.2.1.66
	-	-	4.2.1.131
	-	-	5.2.1.5
	-	-	5.2.1.13
	-	-	5.2.1.14
	-	-	5.3.3.13
	-	-	5.5.1.18
	-	-	5.5.1.19

CHAPTER THREE
Identifying dysregulated metabolic pathways associated with gastric cancer and the effect of *H. pylori* on tumor pathways

3.1 Background

Gastric cancer originates from the stomach mucosa and is the fourth most common cause of cancer-related death in the world. It remains difficult to cure in Western countries, primarily because most patients present with advanced disease [79]. Factors of stomach or gastric cancer can be adenomatous gastric polyps larger than two centimeters, chronic atrophic gastritis, or pernicious anemia. Smoking and consumption of salted, cured or smoked foods can also exacerbate the condition. A genetic component to the cancer is present in approximately 10% of cases and it is mostly found to occur in men over the age of 40. This cancer is most prevalent in Japan, Chile, and Iceland. The diagnosis is done from a biopsy of tissue obtained from gastroscopy. In this procedure, a fiber optic camera is inserted into the stomach via the esophagus and then mucosal samples are scooped or sampled.

Adenocarcinoma is the most common type of stomach cancer. From the histologic point of view there are two main types of adenocarcinoma: intestinal and diffuse. The intestinal subtype is composed of irregular glands that have a "back-to-back" appearance whereas the diffuse subtype is composed of loosely cohesive cells which secrete mucus into the interstitium. Apart from that the intestinal type seems to be more age dependent compared to the diffuse type [80].

Cancer stage is classified according to the TNM system. This system describes the growth of the primary tumor (T), its spread to nearby lymph nodes (N) and the

absence or presence of distant spread known as metastasis (M). In Stage 0, the cancer is just on the inner lining of the stomach. This stomach cancer is treatable in the early stages by endoscopic mucosal resection, or by gastrectomy and lymphadenectomy without a resorting to chemotherapy or radiation. During Stage I the tumor penetrates to the second or third layers of the stomach (IA) or to the nearby lymph nodes (IB). Stage IA is treated by surgery whereas for Stage IB chemotherapy maybe needed. In Stage II, apart from penetrating into the mucosal layers the tumor also spreads to the more distant lymph nodes. The treatment for this is the same as Stage I. Stage III is characterized by beginning of metastasis penetration and by Stage IV the tumor as completely metastasized to distant organs. A cure is very rarely possible at this stages III or IV. The survival rate for each stage of gastric

cancer is highlighted in Table 1.

Stage	5 year observed survival
Stage IA	71%
Stage IB	57%
Stage IIA	46%
Stage IIB	33%
Stage IIIA	20%
Stage IIIB	14%
Stage IIIC	9%
Stage IV	4%

Table 1: Table showing the 5-year survival rates by stage for stomach/gastric cancer treated with surgery. [Cancer.gov]

Spontaneous occurrence of gastric cancer is extremely rare [81]. However, an important development in the epidemiology of gastric carcinoma has been the recognition of the association with *Helicobacter pylori. H. pylori* is a spiral shaped gram negative bacterium that has the ability to grow on gastric epithelial tissue. It grows in the acidic environment of the stomach and converts urea to ammonia creating a viable basic or neutral environment for it to grow [82]. It is found freely in nature and spreads due to contaminated food and water. It is a threat in more developing countries. Signs of infection include feeling bloated, nausea, vomiting, lack of appetite, anorexia and/or unexplained weight loss. Diagnosis is generally by a blood tests that searches for antibodies against *H. pylori*. Some strains have the ability to invade the host cell better due to the presence of "Pathogenicity Islands" (PAIs) resulting in greater persistence [83]. These "islands" are a cluster of approximately 30 genes which can cause activation of transcription factor NF-κB in gastric epithelial cells and are also associated with encoding a type IV secretion [84]. This means that the microbe has the ability to transport virulence proteins right into the host cells. Apart from this the genes can also increase the activity of interleukin-8 (IL-8). IL-8 is the major chemokine causes upregulation of neutrophils to the infection site; thus, increasing inflammation. Due to all these reasons, in 1994, the International Agency for Research on Cancer classified *H. pylori* as a carcinogen depending on its virulence level.

H. pylori is believed to be present in approximately two-thirds of the world's population. It causes more intestinal adenocarcinoma than the diffuse type [85]. However even among infected individuals the susceptibility to cancer is significantly lower. Approximately

10% develop peptic ulcer disease and from there only 1% to 3% develop gastric adenocarcinoma [23]. While the increased chance of gastric cancer for individuals with active *H. pylori* infection has been documented, much remains to be understood about the connection between host-microbe interaction and the onset of gastric cancer.

3.2 Methods

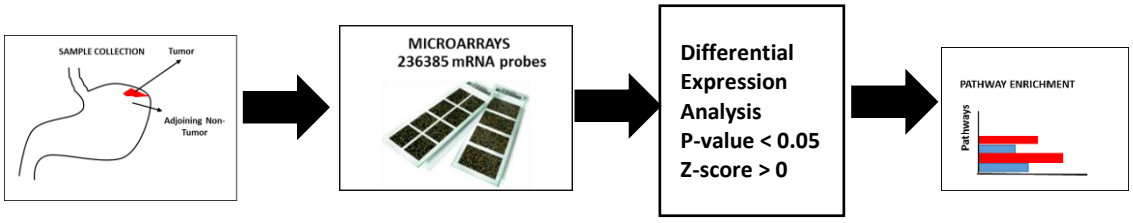
3.2.1 Data

Published data from the GEO [86] (Gene Expression Omnibus) database was used. The first was an Agilent gene expression microarray data for patients affected with gastric cancer - GSE33428 [87]. Cancer is heterogeneous i.e. every person's tumor has a unique growth pattern. In order to reduce this, I selected the same patients for tumor and normal comparison i.e. the normal sample was collected from the adjoining uninfected gastric tissue. This cohort consisted of 27 patients with no information about the ethnicity. The ages ranged from 41- 84 years. A breakdown of the samples can be seen in Table 2 along with the pipeline for analysis in Figure 1.

Sample Type	No. of Samples	No. of Samples used
Gastric cancer tissue	27	27
Gastric non-cancer tissue	27	27

Table 2: Details of dataset1 (GSE33428) used

Figure 1: Pipeline for analysis of GSE33428

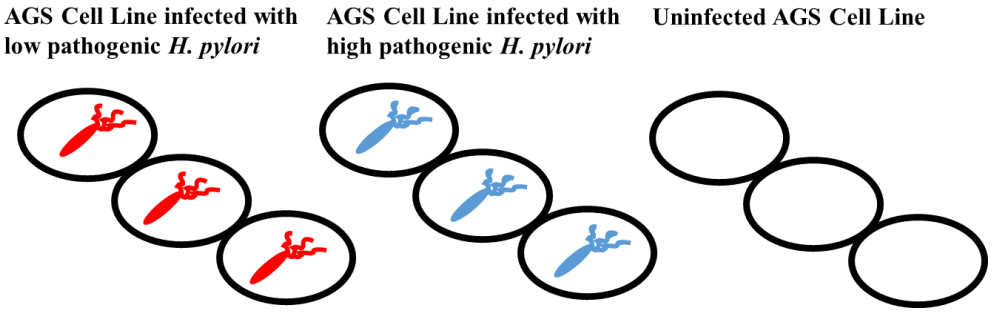


The second was an Affymetrix cell line microarray expression which was specifically derived from gastric cancer adenocarcinoma, the AGS cell line. This cell line was infected with different strains of *H. Pylori* namely high pathogenic and low pathogenic strains - GSE27347. This dataset was chosen as it could predict the difference between the pathogenic and non-pathogenic strains of *H. pylori* as mentioned above in the chapter introduction. The control used in this case was an un-infected cell line. The experiment was done in triplicates for each strain and control. This information is represented in Table 3 and Figure 2.

Samples	Strain A (high pathogenicity)	Strain B (low pathogenicity)	Control
Replicate A	A1 - GSM676126	A2 - GSM676129	A3 - GSM676132
Replicate B	B1 - GSM676127	B2 - GSM676130	B3 - GSM676133
Replicate C	C1 - GSM676128	C2 - GSM676131	C3 - GSM676134

Table 3: Details of the GSE27347 dataset of AGS cell lines infected with different strains of *H. pylori* and control

Figure 2: Cartoon example of Table 3 dataset



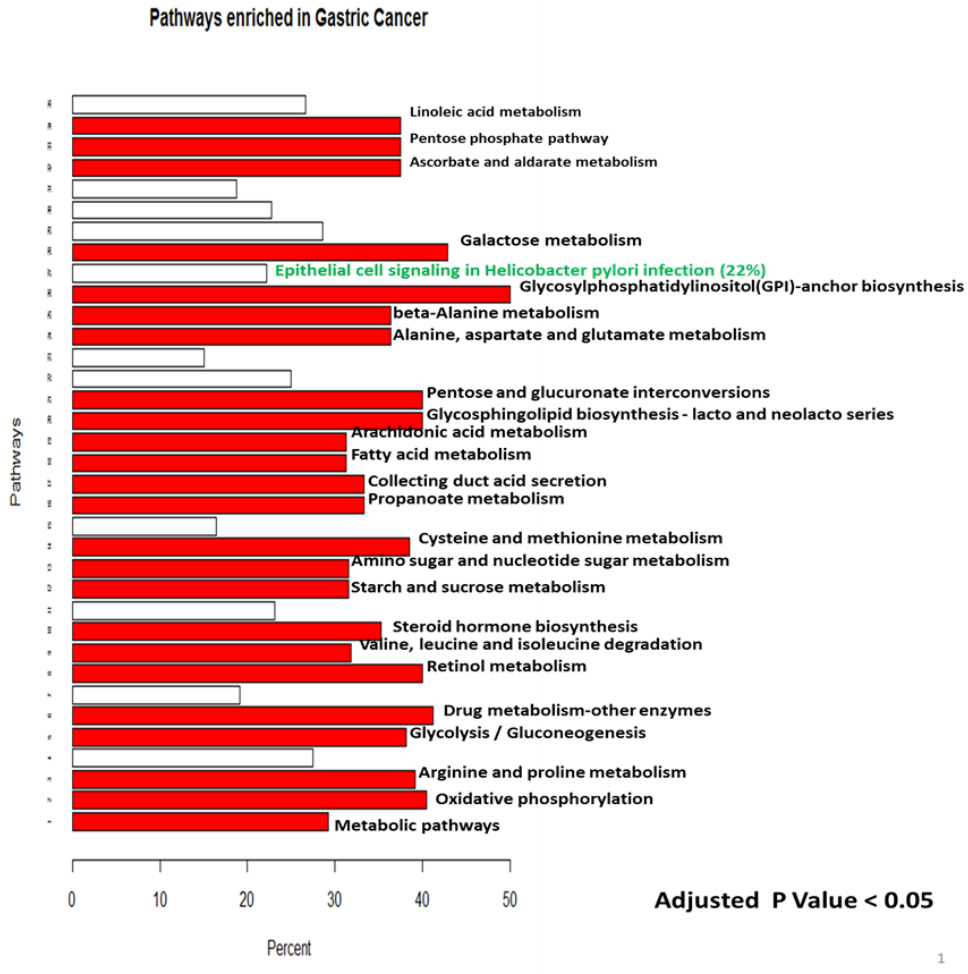
3.2.2 Pre-processing

Dataset one was normalized using the standard RMA suite normalization for microarrays [88]. Following this, a differential analysis was done between tumor and normal samples [89] [90]. I was only interested in the metabolic changes occurring in cancer. So I looked at the genes that mapped to metabolic pathways in KEGG [91] [6]. The rationale for looking at metabolic changes is because these changes are universal in all cancers compared to signaling mechanisms. Also, changing this fundamental pathway provides nutrition to the cancer cells. Recently, microbes have been implicated to have effect on the metabolic pathways too by hijacking them for reproduction. I wanted to see if *H. pylori* has a similar effect that could affect gastric cancer and metastasis.

This reduced the number of probes in the mRNA expression set. P-value of less than 0.05 and a Z-score of greater than 0 was considered as a threshold for significance of the probes. The probes were then mapped to the genes and this produced 156 up-regulated metabolic genes. There were down-regulated genes belonging to metabolic pathway, but they did not fall into the significance category. Pathway enrichment was

done using a hypergeometric test [92]. Around 51 metabolic pathways are enriched as highlighted in Figure 3.

Figure 3: Metabolic pathway enrichments in gastric cancer on comparing tumor and normal samples where the green highlighted pathway shows mRNA dysregulations could involve pathways that can mediate interaction with *H. pylori*



The ones that are highlighted in red are the ones were more than 30% of the genes in the pathways was unregulated. These were mainly pathways that contribute to amino acid and glucose metabolism [93]. This is because the cancer cells utilize the tricarboxylic acid (TCA) cycle to generate energy that is required to feed cancer cells and for fatty acid and phospholipid biosynthesis. Apart from this, another interesting observation was the enrichment in the *H. pylori* signaling pathway for patients who had no occurrence of the microbe. This lead to the first hypothesis that mRNA dysregulations could involve pathways that can mediate interaction with *H. pylori* [Epithelial cell signaling in Helicobacter pylori infection – hsa05120 in KEGG].

In order to substantiate the above hypothesis and determine the causal effect of the microbe in gastric cancer the second dataset was used. The normalization protocol was the same as above. However, with traditional differential expression analysis for comparison between the abnormal samples and the control/normal, thresholds are then used to test significance. The thresholds used are usually p-values or fold changes. In most cases, more than log 2-fold change and p-values of about 0.05 are the usual cut-offs used. However, the drawback with this method is that if there are two abnormal states, though it can capture the total change between them it cannot provide any information as to where the changes came from. So instead of doing a conventional differential expression, I used a new method called the signal-to-noise method. As the name suggests this involves two metrics called ‘signal and ‘noise’. Signal is defined as the difference in gene expression between infected and uninfected sample. Noise is the difference in gene

expression between the uninfected samples. Mathematical representations of the Signal – to – noise-method is highlighted below.

$$Z(i) = \sum_{j,j'} |x(i,j) - x(i,j')|$$

$x =$ expression value in control for probe i, replicate j
i = 1:2511
j = 1:3

$$W(i) = \sum_{j,j'} |y(i,j) - x(i,j')|$$

$x =$ expression value in control for probe i, replicate j
y = expression value in strains for probe i, replicate j
i = 1:2511
j = 1:3

$Z =$ Noise in the data
W = Actual signal

$$\frac{\text{Noise(Z)}}{\text{Signal(W)}} = < 20\% \text{ (i.e. more signal)}$$

3.3 Results

The above new analysis is done by averaging cross all triplicates. The signal-noise method is not only able to capture strain specific differences between high and low pathogenic AGS cell line infected with *H. pylori* but also find the total difference from the control cell line. In other words, it can probe the differences between samples infected with different strains and also tell us which sample it came from. This is clearly demonstrated in Figure 4 and 5. Also performance wise, the new method can detect more dysregulated probes between the samples than the conventional method. As further proof,

I analyzed the data with the new metric as well as the old differential expression pipeline. Comparing the two methods, my analysis with new metric produced 178 probes that are significant in the high pathogenic strain and 63 probes for the low pathogenic strain. Mapping these probes using the NCBI database produced 39 and 12 genes respectively for high and low pathogenic *H. pylori*. This brings the total to 52 genes. On the other hand, the traditional method involving volcano plots seen in Figure 6 only listed 26 differentially expressed probes that corresponds to only 4 genes. This is further demonstrated in Table 4.

	Conventional Differential Expression	**Signal-Noise Method**
Number of probes	26	178 – high pathogenic strain 63 – low pathogenic strain
Total genes	4	52

Table 4: Performance of the Signal-Noise Method compared to Traditional Differential Expression

A heatmap representation of the samples clearly show that the high pathogenic *H. pylori* have a set of genes that are completely up-regulated when compared to low pathogenic strains and control as seen in Figure 7 and Table 5.

GENE LIST				
ABCF1	CHPF	GALNT2	NCSTN	SMC3
ACVR2B	CHST14	GALNT3	NDST1	SMNDC1
ANAPC5	CHST3	GALNT4	NEU1	SRP14
APH1A	COG1	GDI2	NONO	ST6GAL1
ARF1	COG4	GNPNAT1	NOTCH1	STARD7
ASAH1	COG6	HINT1	NPL	STS
ATP6V0B	COL9A2	HNRPM	PAPSS2	TAF10
B3GALT6	DGCR2	HPSE2	PARK7	TARDBP
B3GNT5	DLL3	HS2ST1	PODXL2	TCEB2
B4GALT7	DPM1	HS3ST1	POLR2I	TPST1
BMP4	DULLARD	HSP90AB1	PSEN2	TPST2
BMP7	EEF2	IGF2R	PSMB2	UGP2
C1GALT1C1	EIF3D	IL6ST	RFNG	ZNF347
C1orf103	ERH	JAG2	RPL24	
C2orf24	FAU	JTB	RPL6	
CANX	FNTA	JUND	RPS11	
CCL22	GAL3ST2	KARS	SEP	
CCL5	GALNT10	KLRF1	SGSH	
CFL1	GALNT14	LGALS12	SLC35B2	

Table 5: List of 89 genes that are dysregulated between high pathogenic vs low pathogenic and control

The next step involved performing a T-Test to identify significant probes between a particular strain and the control. This was done in order to find processes that might be different between high and low pathogenic strains of *H. pylori* providing a better insight

into mechanistic behavior of the microbes that might explain its interaction with gastric cancer. GO Term enrichment with clustering with average linkage shows that the high pathogenic *H. pylori* strains have a greater impact or enrichment for metabolic genes (p-value < 0.05) than they low pathogenic counterpart seen in Figure 8. This provides an idea as to how the high pathogenic strains of the bacteria can be virulent and cause more malignancy. That means the high pathogenic microbe might allow the cancer cells to hijack the metabolic pathways that can cause a decrease in healthy cells and progression and differentiation of tumor cells. To prove the hypothesis further and establish metabolic pathways that the high pathogenic *H. pylori* can regulate for tumorogenictiy; pathway enrichment was done using GSEA [94]. The pathway enrichments are in Figure 9.

A closer look at the pathway enrichments of the two strains we see classic cancer signaling pathways dysregulated. These would include the Jak-STAT pathway, Cytokine Receptor pathway, VEGF and mTOR Signaling. For example, the activation of VEGF causes neovascularization, facilitating in production of more cancer cells [95]. However, apart from that I noticed enrichment in the purine metabolism pathway in the high pathogenic strain. It has been recently known that purinergic signaling plays an important part in inflammation and cancer [96]. This further proves that the high pathogenic *H. pylori* can exacerbate the cancer compared to the low pathogenic strain. The mechanism of inflammation is that this causes the release of ATP which is a part of the signaling process. The ATP has 2 receptors P2X and P2Y. There are seven P2X

receptors and eight P2Y receptors [97]. The P2Y receptor is the one that is most affected by cancer.

Out of the eight subtypes of P2Y, five purinoceptors have a greater effect on cancer by affecting several key processes. $P2Y_1$ and $P2Y_2$ receptors might affect the rate of cell proliferation. $P2X_5$ and $P2Y_{11}$ receptor activation might causes changes in cell cycle and the $P2X_7$ receptor activates the apoptosis. It is known in gastric cancer; that apoptosis could be mediated with ATP and adenosine [98] and *Helicobacter pylori* could contribute to the progression of gastric carcinoma, perhaps by regulation of H,K-ATPase [99]. However, more work would need to be done to show the association of the microbe with a particular purine receptor.

Figure 4: Calculating differences in the high pathogenic samples compared to Control

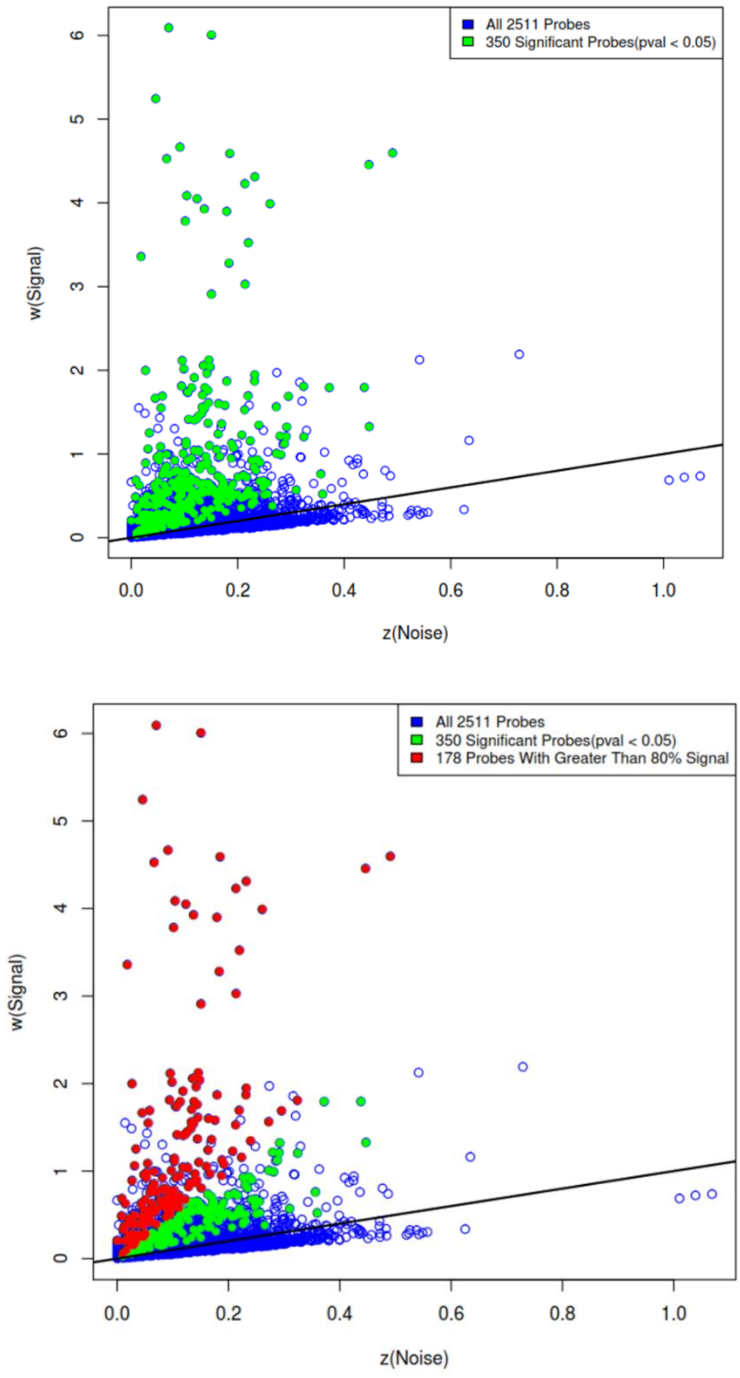


Figure 5: Calculating differences in the low pathogenic samples compared to Control

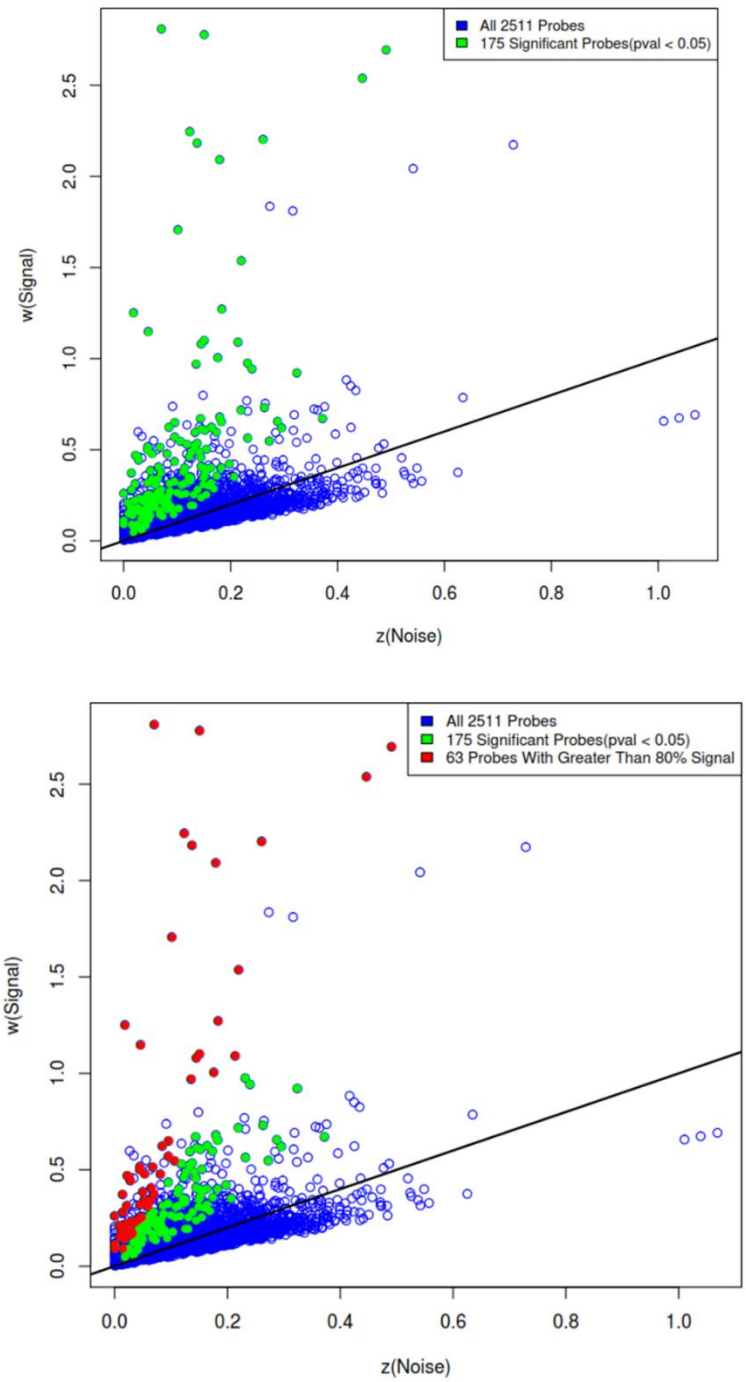


Figure 6: Volcano Plot for genes that are differentially expressed between strains and control

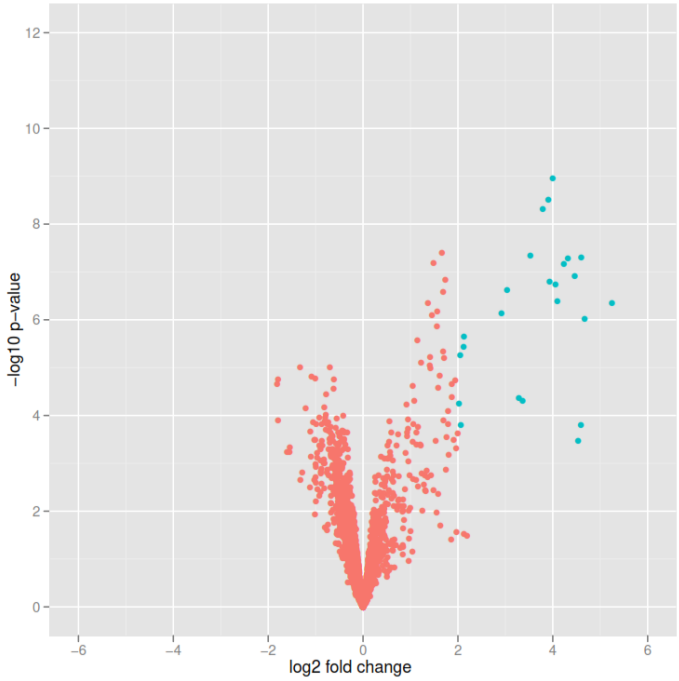


Figure 7: Heatmap of different strains shows dysregulation of 89 genes between high pathogenic vs low pathogenic and control

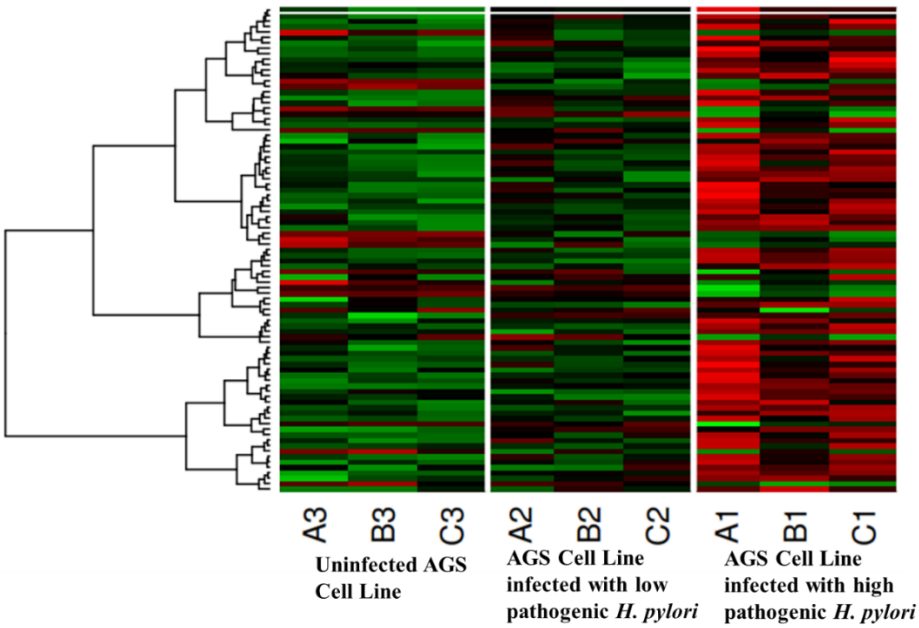
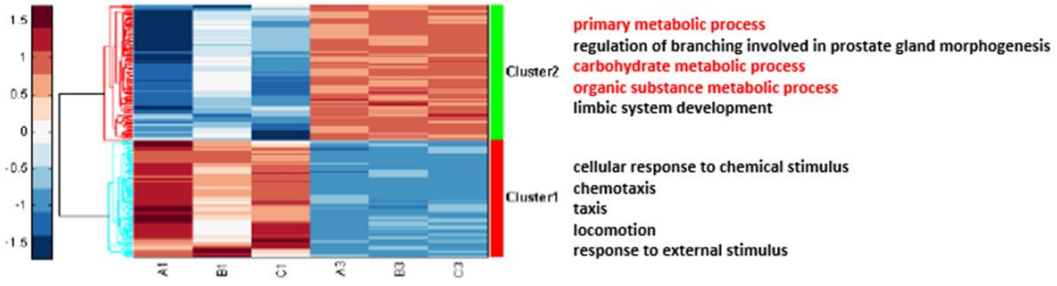


Figure 8: (A) Go Term enrichment in the high pathogenic strains
(B) Go Term enrichment in the low pathogenic strains;
where the metabolic pathways are colored in red



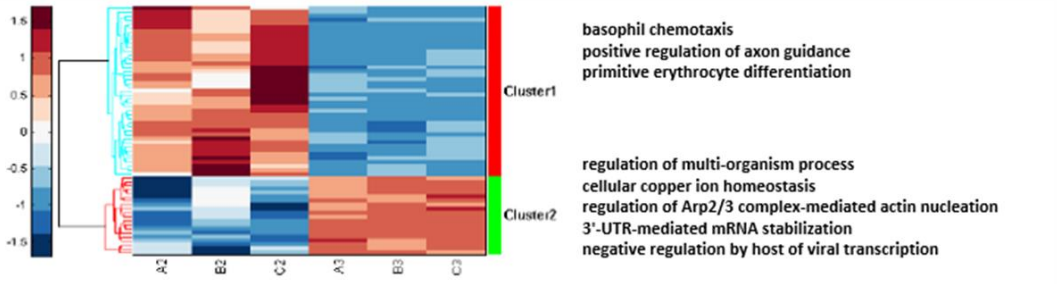
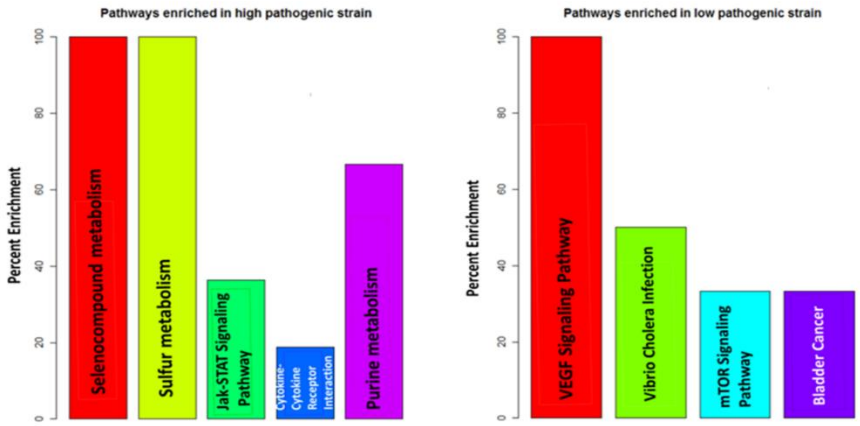


Figure 9: Pathway enrichment with GSEA



3.4 Discussion

Thus in this chapter I have tried to elucidate the mechanism of association of the gram negative bacteria *H. pylori* to gastric cancer. Novel ways of analyzing differential gene expression made it clear that different strains of the microbe affect the cancer in different ways. This also further validates the claim as to how many people who have the microbe remain asymptomatic to it depending on the microbe type.

Virulent strains of the bacteria have a greater effect on metabolic pathways than the less virulent strains. This makes the former strains more tumorigenic. Pathway analysis provides a possible method of how the metabolic dysregulations of the high pathogenic strain can cause metastasis and affect the host. One such pathway that is affected in the high pathogenic and virulent strains is purine metabolism. As mentioned above, this pathway has a direct association with inflammation and P2Y receptors. Several researches into this pathway and its receptors have shown that many of these are implicated with cancer. Thus, this sort of analysis provides us with a better understanding of different *H. pylori* bugs and methods to treat them.

CHAPTER FOUR: General Conclusions and Future Directions

In this thesis, I have used high throughput 16S sequencing and microarray data to predict how microbes interact with the environment. The results from the chapters prove that:

- For the order of colonization and context of enzymes the microbes in different stages strongly depend on the microbes in the previous step.
- The diet plays an important part in determining the organisms and their diversity. Using network expansion approaches I can predict the organisms that would best represent the diet profile.
- On a microscale, looking at the interaction of *H. pylori* and gastric cancer, I can also conclude that mRNA dysregulations in the human host can mediate interactions with the microbe. Also, I see that based on reviewed literature not all strains of the bacteria affect gastric cancer. On comparison of the bacteria with different strains and pathogenicity levels the highly pathogenic ones have a greater effect on the gastric cells by severely affecting inflammation pathways.

However, as mentioned before, there are several methods that can be pursued to improve the results. The first method would be to test the microbiome pipeline and hypothesis for a larger sample size for microbiome data. The second method is to improve the network expansion by leveraging genome scale metabolic models [100]. These models have information about the enzyme stoichiometry of the organisms. Apart from that, it contains the biomass reaction which states the necessary components for growth

of the organism. This biomass reaction could be used to curate the amounts of the different diets used as input for the network expansion. A vast number of manually curated gut microbiome metabolic reconstructions were recently published [101]. These would be a good source for thresholding the input diet metabolites.

Lastly experiments could be set up where we can verify the network expansion algorithm and also the *H. pylori* hypothesis. For the former, *in vivo* experiments can be devised where the germ-free mice are fed different amounts of the diet and we can measure the different amounts of compounds produced by them, for instance short chain fatty acids, and then correlate these compounds back to the organisms that produced them in the diet. This way the changing amounts of diets could predict which organisms are the top drivers of metabolism in the diet. As for the *H. pylori* and the hypothesis that inflammation can be acerbated by the bacteria by upregulation of purine metabolism; gene knockout experiments can be done. The genes in the pathway can be knocked out to see how that affects the tumor cell and progression.

LIST OF JOURNAL ABBREVIATIONS

Appl. Environ. Microbiol.	Applied and Environmental Microbiology
Am. J. Physiol.	American Journal of Physiology
BMC Med. Genomics	BMC Medical Genomics
Biochem. Pharmacol.	Biochemical Pharmacology
Br. J. Cancer	British Journal of Cancer
Crit. Rev. Food Sci. Nutr.	Critical Reviews in Food Science and Nutrition
Curr. Opin. Biotechnol.	Current Opinion in Biotechnology
Curr. Opin. Gastroenterol.	Current Opinion in Gastroentology
Eur. J. Clin. Nutr.	European Journal of Clinical Informatics
Front. Microbiol.	Frontiers in Microbiology
Genome Inform.	Genome Informatics
J. Agric. Food Chem.	Journal of Agriculture and Food Chemistry
J Clin Gastroenterol	Journal of Clinical Gastroenterology
J. Physiol.	Journal of Physiology
J. Mol. Evol.	Journal of Molecular Evolution
J. Natl. Cancer Inst.	Journal of National Camcer Insitute
J. Nutr.	Journal of Nutrition
J. Pediatr.	Journal of Pediatrics
J. Pediatr. Gastroenterol. Nutr.	Journal of Pedriatric Gastroentology and Nutrition
J. Zhejiang Univ. Sci.	Journal of Zhejiang University
Lipids Health Dis.	Lipids in Health and Disease

Microb. Ecol. Health Dis.	Microbial Ecology in Health and Disease
Nat. Biotechnol.	Nature Biotechnology
Nat. Chem. Biol.	Nature Chemical Biology
Nat. Immunol.	Nature Immunology
Nat. Methods	Nature Methods
Nat. Rev. Genet.	Nature Reviews Genetics
Nat. Rev. Microbiol.	Nature Reviews Microbiology
Nucleic Acids Res.	Nucleic Acids Research
PLoS Biol.	PLos Biology
PLoS Comput. Biol.	PLoS Computational Biology
Proc. Natl. Acad. Sci	Proceedings of National Academy of Sciences
Sci. Rep.	Scientific Reports

BIBLIOGRAPHY

1. Bäckhed, F. *et al.* Defining a healthy human gut microbiome: current concepts, future directions, and clinical applications. *Cell Host Microbe* **12,** 611–622 (2012).
2. Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L. & Gordon, J. I. Human nutrition, the gut microbiome and the immune system. *Nature* **474,** 327–336 (2011).
3. Khoruts, a, Dicksved, J., Jansson, J. K. & Sadowsky, M. J. Changes in the composition of the human fecal microbiome after bacteriotherapy for recurrent Clostridium difficile-associated diarrhea. *J Clin Gastroenterol* **44,** 354–360 (2010).
4. Waldor, M. K. *et al.* Where Next for Microbiome Research? *PLoS Biol.* **13,** (2015).
5. Franzosa, E. a. *et al.* Sequencing and beyond: integrating molecular 'omics' for microbial community profiling. *Nat. Rev. Microbiol.* **13,** 360–72 (2015).
6. Ogata, H. *et al.* KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **27,** 29–34 (1999).
7. Finn, R. D. *et al.* Pfam: The protein families database. *Nucleic Acids Research* **42,** (2014).
8. EMBL, SIB Swiss Institute of Bioinformatics & Protein Information Resource (PIR). UniProt. in *Nucleic acids research* 41: D43-D47 (2013).
9. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution.

Nucleic Acids Res. **28,** 33–36 (2000).

10. Yeung, N., Cline, M. S., Kuchinsky, A., Smoot, M. E. & Bader, G. D. Exploring biological networks with cytoscape software. *Current Protocols in Bioinformatics* (2008).
11. Karp, P. D., Riley, M., Paley, S. M. & Pellegrini-Toole, A. The MetaCyc Database. *Nucleic Acids Res.* **30,** 59–61 (2002).
12. Larsen, P. E. & Dai, Y. Metabolome of human gut microbiome is predictive of host dysbiosis. *Gigascience* **4,** 42 (2015).
13. Larsen, P. E., Field, D. & Gilbert, J. A. Predicting bacterial community assemblages using an artificial neural network approach. *Nat. Methods* **9,** 621–5 (2012).
14. Borenstein, E., Kupiec, M., Feldman, M. W. & Ruppin, E. Large-scale reconstruction and phylogenetic analysis of metabolic environments. *Proc. Natl. Acad. Sci.* **105,** 14482–14487 (2008).
15. Handorf, T., Ebenhöh, O. & Heinrich, R. Expanding metabolic networks: Scopes of compounds, robustness, and evolution. *J. Mol. Evol.* **61,** 498–512 (2005).
16. Kolenbrander, P. E., Palmer, R. J., Periasamy, S. & Jakubovics, N. S. Oral multispecies biofilm development and the key role of cell-cell distance. *Nat. Rev. Microbiol.* **8,** 471–480 (2010).
17. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457,** 480–484 (2009).
18. Houghteling, P. D. & Walker, W. A. Why Is Initial Bacterial Colonization of the

Intestine Important to Infants' and Children's Health? *J. Pediatr. Gastroenterol. Nutr.* **60,** 294–307 (2015).

19. Zomorrodi, A. R. & Maranas, C. D. OptCom: A multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS Comput. Biol.* **8,** (2012).

20. Greenblum, S., Chiu, H.-C., Levy, R., Carr, R. & Borenstein, E. Towards a predictive systems-level model of the human microbiome: progress, challenges, and opportunities. *Curr. Opin. Biotechnol.* **24,** 810–20 (2013).

21. Chu, J. *et al.* Discovery of MRSA active antibiotics using primary sequence from the human microbiome. *Nat. Chem. Biol.* (2016). doi:10.1038/nchembio.2207

22. Scott, D., Weeks, D., Melchers, K. & Sachs, G. The life and death of Helicobacter pylori. *Gut* **43 Suppl 1,** S56–S60 (1998).

23. Wroblewski, L. E., Peek, R. M. & Wilson, K. T. Helicobacter pylori and gastric cancer: Factors that modulate disease risk. *Clinical Microbiology Reviews* **23,** 713–739 (2010).

24. Alison Abbot. Scientists bust myth that our bodies have more bacteria than human cells. *Nature* (2016). doi:10.1038

25. Johnson, C. L. & Versalovic, J. The human microbiome and its potential importance to pediatrics. *Pediatrics* **129,** 950–60 (2012).

26. Degnan, P. H., Taga, M. E. & Goodman, A. L. Vitamin B12 as a modulator of gut microbial ecology. *Cell Metabolism* **20,** 769–778 (2014).

27. Hehemann, J.-H., Kelly, A. G., Pudlo, N. a, Martens, E. C. & Boraston, A. B.

Bacteria of the human gut microbiome catabolize red seaweed glycans with carbohydrate-active enzyme updates from extrinsic microbes. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 19786–91 (2012).

28. Rodríguez, J. M. *et al.* The composition of the gut microbiota throughout life, with an emphasis on early life. *Microb. Ecol. Health Dis.* **26,** 26050 (2015).
29. Marcobal, A. *et al.* Metabolome progression during early gut microbial colonization of gnotobiotic mice. *Sci. Rep.* **5,** 11589 (2015).
30. Jenkinson, H. F. & Lamont, R. J. Oral microbial communities in sickness and in health. *Trends in Microbiology* **13,** 589–595 (2005).
31. Favier, C. F., Vaughan, E. E., De Vos, W. M. & Akkermans, A. D. L. Molecular monitoring of succession of bacterial communities in human neonates. *Appl. Environ. Microbiol.* **68,** 219–226 (2002).
32. Zijnge, V. *et al.* Oral biofilm architecture on natural teeth. *PLoS One* **5,** (2010).
33. Lee, S. M. *et al.* Bacterial colonization factors control specificity and stability of the gut microbiota. *Nature* **501,** 426–429 (2013).
34. Martens, E. C., Chiang, H. C. & Gordon, J. I. Mucosal Glycan Foraging Enhances Fitness and Transmission of a Saccharolytic Human Gut Bacterial Symbiont. *Cell Host Microbe* **4,** 447–457 (2008).
35. Sonnenburg, J. L. *et al.* Glycan foraging in vivo by an intestine-adapted bacterial symbiont. *Science* **307,** 1955–9 (2005).
36. Baxter, N. T., Zackular, J. P., Chen, G. Y. & Schloss, P. D. Structure of the gut microbiome following colonization with human feces determines colonic tumor

burden. *Microbiome* **2,** 20 (2014).

37. Benítez-Páez, A., Belda-Ferre, P., Simón-Soro, A. & Mira, A. Microbiota diversity and gene expression dynamics in human oral biofilms. *BMC Genomics* **15,** 311 (2014).

38. Mueller, N. T., Bakacs, E., Combellick, J., Grigoryan, Z. & Dominguez-Bello, M. G. The infant microbiome development: Mom matters. *Trends in Molecular Medicine* **21,** 109–117 (2015).

39. Pearl D. Houghteling & W. Allan Walker. Why is initial bacterial colonization of the intestine important to the infant's and child's health? *J Pediatr Gastroenterol Nutr.* (2015).

40. Devaraj, S., Hemarajata, P. & Versalovic, J. The human gut microbiome and body metabolism: Implications for obesity and diabetes. *Clinical Chemistry* **59,** 617–628 (2013).

41. McKenney, E. A., Rodrigo, A. & Yoder, A. D. Patterns of gut bacterial colonization in three primate species. *PLoS One* **10,** (2015).

42. Mazumdar, V., Amar, S. & SegrÃ¨, D. Metabolic Proximity in the Order of Colonization of a Microbial Community. *PLoS One* **8,** e77617 (2013).

43. Koenig, J. E. *et al.* Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl,** 4578–4585 (2011).

44. Faith, J. J. *et al.* Predicting a human gut microbiota's response to diet in gnotobiotic mice. *Science* **333,** 101–4 (2011).

45. Sonnenburg, E. D. *et al.* Diet-induced extinctions in the gut microbiota compound

over generations. *Nature* **529,** 212–215 (2016).

46. David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505,** 559–63 (2014).

47. Turnbaugh, P. J., Bäckhed, F., Fulton, L. & Gordon, J. I. Diet-Induced Obesity Is Linked to Marked but Reversible Alterations in the Mouse Distal Gut Microbiome. *Cell Host Microbe* **3,** 213–223 (2008).

48. Turnbaugh, P. J. & Gordon, J. I. The core gut microbiome, energy balance and obesity. *J. Physiol.* **587,** 4153–8 (2009).

49. Marcobal, A. *et al.* Consumption of human milk oligosaccharides by gut-related microbes. *J. Agric. Food Chem.* **58,** 5334–5340 (2010).

50. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12,** 56–68 (2011).

51. Leinonen, R., Sugawara, H. & Shumway, M. The sequence read archive. *Nucleic Acids Res.* **39,** (2011).

52. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7,** 335–6 (2010).

53. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–10 (1990).

54. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72,** 5069–5072 (2006).

55. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 - Approximately maximum-

likelihood trees for large alignments. *PLoS One* **5,** (2010).

56. Ballard, O. & Morrow, A. L. Human Milk Composition. Nutrients and Bioactive Factors. *Pediatric Clinics of North America* **60,** 49–74 (2013).

57. Andreas, N. J., Kampmann, B. & Mehring Le-Doare, K. Human breast milk: A review on its composition and bioactivity. *Early Human Development* **91,** 629–635 (2015).

58. Koehler, P. & Wieser, H. in *Handbook on Sourdough Biotechnology* (eds. Gobbetti, M. & Gänzle, M.) 245–264 (2013). doi:10.1007/978-1-4614-5425-0

59. Lennox, A., Olson, a & Gay, C. National diet and nutrition survey. *Headl. results from Years* **2,** 2009–2009. (2011).

60. Haug, A., Høstmark, A. T. & Harstad, O. M. Bovine milk in human nutrition--a review. *Lipids Health Dis.* **6,** 25 (2007).

61. Eskin, N. A. M. *Biochemistry of Foods: Second Edition*. *Biochemistry of Foods: Second Edition* (2012).

62. United States Department of Agriculture - USDA. Available at: https://ndb.nal.usda.gov/ndb/search/list?fgcd=Branded+Food+Products+Database&ds=Branded+Food+Products.

63. Smilowitz, J. T., Sullivan, A. O. Õ., Barile, D., German, J. B. & Lo, B. The Human Milk Metabolome Reveals Diverse Oligosaccharide Profiles. *J. Nutr.* **143,** 1709–1718 (2013).

64. Calvo, M. M. Lutein: a valuable ingredient of fruit and vegetables. *Crit. Rev. Food Sci. Nutr.* **45,** 671–696 (2005).

65. Tester, R. F., Karkalas, J. & Qi, X. Starch - Composition, fine structure and architecture. *Journal of Cereal Science* **39,** 151–165 (2004).

66. Jewell, V. C., Mayes, C. B. D., Tubman, T. R. J., Northrop-Clewes, C. a & Thurnham, D. I. A comparison of lutein and zeaxanthin concentrations in formula and human milk samples from Northern Ireland mothers. *Eur. J. Clin. Nutr.* **58,** 90–97 (2004).

67. Mock, D. M., Mock, N. I. & Stratton, S. L. Concentrations of biotin metabolites in human milk. *J. Pediatr.* **131,** 456–8 (1997).

68. Sindayikengera, S. & Xia, W. Nutritional evaluation of caseins and whey proteins and their hydrolysates from Protamex. *J. Zhejiang Univ. Sci. B* **7,** 90–8 (2006).

69. Raymond, J. & Segrè, D. The effect of oxygen on biochemical networks and the evolution of complex life. *Science (80-.).* **311,** 1764–1767 (2006).

70. Ebenhöh, O., Handorf, T. & Heinrich, R. Structural analysis of expanding metabolic networks. *Genome Inform.* **15,** 35–45 (2004).

71. Thomas, F., Hehemann, J. H., Rebuffet, E., Czjzek, M. & Michel, G. Environmental and gut Bacteroidetes: The food connection. *Front. Microbiol.* **2,** (2011).

72. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22,** 1658–1659 (2006).

73. Hebbeln, P., Rodionov, D. A., Alfandega, A. & Eitinger, T. Biotin uptake in prokaryotes by solute transporters with an optional ATP-binding cassette-containing module. *Proc Natl Acad Sci U S A* **104,** 2909–2914 (2007).

74. Conlon, M. & Bird, A. The Impact of Diet and Lifestyle on Gut Microbiota and Human Health. *Nutrients* **7,** 17–44 (2014).

75. Hieken, T. J. *et al.* The Microbiome of Aseptically Collected Human Breast Tissue in Benign and Malignant Disease. *Sci. Rep.* **6,** 30751 (2016).

76. Jost, T., Lacroix, C., Braegger, C. P. & Chassard, C. New Insights in Gut Microbiota Establishment in Healthy Breast Fed Neonates. *PLoS One* **7,** (2012).

77. Zhu, Y. *et al.* Meat, dairy and plant proteins alter bacterial composition of rat gut bacteria. *Sci. Rep.* **5,** 15220 (2015).

78. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is Flux Balance Analysis ? *Nat. Biotechnol.* **28,** 245–248 (2010).

79. National Cancer Institute - NCI. Available at: http://www.cancer.gov/.

80. Lundegårdh, G. *et al.* Intestinal and diffuse types of gastric cancer: secular trends in Sweden since 1951. *Br. J. Cancer* **64,** 1182–1186 (1991).

81. Pritchard, D. M. & Crabtree, J. E. Helicobacter pylori and gastric cancer. *Curr. Opin. Gastroenterol.* **22,** 620–625 (2006).

82. Peek, R. M. & Crabtree, J. E. Helicobacter infection and gastric neoplasia. *Journal of Pathology* **208,** 233–248 (2006).

83. Stein, M., Rappuoli, R. & Covacci, A. in *Helicobacter pylori: Physiology and Genetics* (eds. Mobley, H. L. T., Mendz, G. L. & Hazell, S. L.) (2001).

84. Viala, J. *et al.* Nod1 responds to peptidoglycan delivered by the Helicobacter pylori cag pathogenicity island. *Nat. Immunol.* **5,** 1166–1174 (2004).

85. Parsonnet, J. *et al.* Helicobacter pylori infection in intestinal- and diffuse-type

gastric adenocarcinomas. *J. Natl. Cancer Inst.* **83,** 640–643 (1991).

86. Edgar, R., Domrachev, M. & Lash, A. E. *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository*. *Nucleic Acids Res* **30,** 207–210 (2002).

87. Cheng, L. *et al.* Identification of genes with a correlation between copy number and expression in gastric cancer. *BMC Med. Genomics* **5,** 14 (2012).

88. Irizarry, R. A. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31,** e15 (2003).

89. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47 (2015).

90. Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. Affy - Analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20,** 307–315 (2004).

91. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44,** D457–D462 (2016).

92. Rivals, I., Personnaz, L., Taing, L. & Potier, M. C. Enrichment or depletion of a GO category within a class of genes: Which test? *Bioinformatics* **23,** 401–407 (2007).

93. Benjamin, D. I., Cravatt, B. F. & Nomura, D. K. Global profiling strategies for mapping dysregulated metabolic pathways in cancer. *Cell Metabolism* **16,** 565–577 (2012).

94. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach

for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 15545–50 (2005).

95. Kieran, M. W., Kalluri, R. & Cho, Y. J. The VEGF pathway in cancer and disease: responses, resistance, and the path forward. *Cold Spring Harb Perspect Med* **2,** a006593 (2012).

96. White, N. & Burnstock, G. P2 receptors and cancer. *Trends in Pharmacological Sciences* **27,** 211–217 (2006).

97. Burnstock, G. & Di Virgilio, F. Purinergic signalling and cancer. *Purinergic Signalling* **9,** 491–540 (2013).

98. Saitoh, M. *et al.* Adenosine induces apoptosis in the human gastric cancer cells via an intrinsic pathway relevant to activation of AMP-activated protein kinase. *Biochem. Pharmacol.* **67,** 2005–2011 (2004).

99. Saha, A., Hammond, C. E., Gooz, M. & Smolka, A. J. The role of Sp1 in IL-1beta and H. pylori-mediated regulation of H, K-ATPase gene transcription. *Am. J. Physiol. Gastrointest. Liver Physiol.* **295,** G977–G986. (2008).

100. Henry, C. S. *et al.* High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.* **28,** 977–982 (2010).

101. Magnúsdóttir, S. *et al.* Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat. Biotechnol.* (2016). doi:10.1038/nbt.3703

CURRICULUM VITAE

AMRITA KAR

Contact Information

20 Exchange Place
Apt.803
NYC, NY 10005
akar@bu.edu
571-294-7877

Education

2010-2017 Boston University, Boston, MA
Segre Lab
Ph.D. Bioinformatics

2013 Boston University, Boston, MA
M.S. Bioinformatics

2006-2010 West Bengal University of Technology
B.Tech Information Technology

Publications

William R. Harcombe, William J. Riehl, Ilija Dukovski, Brian R. Granger, Alex Betts, Alex H. Lang,
Gracia Bonilla,Amrita Kar, et al. (2014) Metabolic Resource Allocation in Individual Microbes
Determines Ecosystem Interactions and Spatial Dynamics. Cell Reports. Volume 7, Issue 4, 1104 -1115

Research Experience

PhD Candidate, Boston University 2010 – present

- Studied bacterial growth over time and space. Developed aspects of a software that implements this using diffusion algorithms, flux balance analysis and JAVA. [Reference: Publications]
- Analyzed microarray datasets to determine differentially expressed genes in gastric cancer cell line data infected with different pathogenic strains of H. pylori.

- Performed pathway enrichments on gastric cancer samples with H. pylori v/s only tumor samples and analyzed the dysregulations with H. pylori types.
- Currently studying the order of colonization and the effects of diet in the infant human gut using
  16S sequencing data obtained from the SRA (NCBI) database.
- Created custom pipelines of scripts (Perl, Python, bash) to analyze large numbers of clinical 16S gut microbial samples in an automated fashion.
- Performed statistical analyses on the 16S data to determine diet correlation with microbiome and colonization importance.
- Mentoring and supervising a M.S. bioinformatics student.

Group Project, Boston University 2010 – 2011

- Worked in a collaborative group consisting of four bioinformatics PhD students.
- Mined and analyzed microarray data aimed at identifying genes that have been differentially expressed in the Ebola Virus for primates.
- Advised experimentalists on experiments suitable to validate the computational results.

Undergraduate Research, West Bengal University of Technology 2006 - 2010

- Developed advanced stochastic clustering algorithms for efficient gene expression using clustering algorithms and neural networks (2009 - 2010).

Selected Abstracts and Conference Presentations

Amrita Kar, Gyan Bhanot and Daniel Segre (2014) Identifying dysregulated metabolic pathways associated with gastric cancer and effect of H. pylori on tumor growth. Intelligent Systems for
Molecular Biology (ISMB). Boston MA, USA.

Amrita Kar, William R. Harcombe, William J. Riehl, Ilija Dukovski, Brian R. Granger, Alex Betts, Alex H. Lang, Gracia Bonilla, Nicholas Leiby, Pankaj Mehta, Christopher J. Marx and Daniel Segre (2012) COMETS : A platform for spatio-temporal stoichiometric modeling of metabolism in microbial ecosystems. Boston Bacterial Meeting. Boston, MA, USA.

Amrita Kar, William R. Harcombe, William J. Riehl, Ilija Dukovski, Brian R. Granger, Alex Betts, Alex H. Lang, Gracia Bonilla, Nicholas Leiby, Pankaj Mehta, Christopher J. Marx and Daniel Segre (2012) COMETS : A platform for spatio-temporal stoichiometric modeling of metabolism in microbial ecosystems. Genomic Science meeting of the Department of Energy (DOE). Virginia, VA,
USA.

Amrita Kar, Evan Appleton, Teresa Wang, Viktor Vassilev (2011) Comparison of global transcriptional host responses across non-human primates infected with anthrax, poxvirus and filoviruses. 11th Annual International Workshop on Bioinformatics and Systems Biology (IWBSB). Berlin, Germany.